Modèle de langue pour l'ordonnancement conjoint d'entités pertinentes dans un réseau d'informations hétérogènes

Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine, Wahiba Bahsoun

Modèle de langue pour l'ordonnancement conjoint d'entités pertinentes dans un réseau d'informations hétérogènes

Laure Soulier, Lamjed Ben Jabeur
Lynda Tamine, Wahiba Bahsoun

IRIT, Université Paul Sabatier
118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9
soulier, jabeur, tamine, wbahsoun@irit.fr

RÉSUMÉ. *Dans ce papier, nous proposons un nouveau modèle, appelé BibRank, ayant pour objectif d'ordonnancer conjointement des ressources hétérogènes, documents et auteurs, d'un réseau bibliographique selon leur degré de pertinence vis-à-vis d'une requête. Ce modèle utilise le principe de propagation des scores des entités en considérant à la fois la structure du réseau et le sujet de la requête. De plus, ce modèle introduit deux indicateurs de proximité thématique entre entités connectées suivant le type des entités reliées. Pour les relations entre entités homogènes, cet indicateur détecte les citations marginales tandis que pour les relations entre entités hétérogènes, il utilise deux sources d'évidence : le sujet du document et l'expertise de l'auteur. Des expérimentations, menées en utilisant le réseau bibliographique CiteSeerX, montrent l'efficacité du modèle d'ordonnancement proposé.*

ABSTRACT. *This paper proposes a novel model for co-ranking heterogeneous entities, namely authors and documents, within a bibliographic network according to their relevance to a query. This model uses a score propagation algorithm taking into account the graph structure and the query topic. Furthermore, this model integrates two content-based indicators in order to estimate the topical relatedness between connected entities. For relationships between homogeneous entities, this indicator detects marginal citations while it connects two evidence sources for relationships between heterogeneous entities : the document topic and the author's knowledge. Experiments on the CiteSeerX dataset show the effectiveness of our model.*

MOTS-CLÉS : *Multi-ordonnancement, réseau bibliographique, réseau hétérogène et homogène*

KEYWORDS: *Multi-ranking, bibliographic network, heterogeneous and homogeneous network*

1. Introduction

De nombreux domaines, tels que la biologie (Roy *et al.*, 2008) ou les transports (Emmerink, 1993), utilisent les réseaux d'informations caractérisés par un nombre important d'acteurs, appelés aussi entités, et de relations entre ces derniers. Les réseaux hypertextes constituent un exemple de réseaux homogènes composés de documents connectés par des liens hypertextes (Page *et al.*, 1999). Avec l'avènement du web 2.0, des réseaux de plus en plus complexes apparaissent incluant, par exemple, des réseaux sociaux, tels que Twitter[1] (Agarwal *et al.*, 2008), des réseaux multimedia tels que Flickr[2] (Rorissa, 2010) ou encore des réseaux de professionnels comme LinkedIn[3]. Ces derniers réseaux sont composés de différents types d'entités, telles que les individus, entreprises, groupes ou encore productions documentaires, connectées par des relations comme *publier*, *être affilié*, *travailler*. Les réseaux multimédia sont, quant à eux, composés d'entités dénotant principalement des individus et des documents sonores, graphiques ou vidéos. Ces entités sont connectées par les relations suivantes : *produire*, *annoter*, *visualiser*, ...

Dans nos travaux, nous nous intéressons à une tâche de recherche d'information (RI) dans les réseaux d'informations hétérogènes. L'objectif est de formaliser, à partir d'un besoin utilisateur modélisé par une requête, une fonction d'estimation de la pertinence des entités en considérant la structure du réseau, permettant ainsi de les ordonnancer. Le défi majeur introduit par la RI dans les réseaux d'informations hétérogènes est de proposer un ordonnancement conjoint de plusieurs types d'entités en fonction de la structure du réseau. Dans le contexte de ce travail, nous nous intéressons spécifiquement à un réseau d'informations particulier : le réseau d'informations bibliographiques. Ce réseau est composé d'auteurs et de documents scientifiques connectés par des relations de citation entre auteurs ou entre documents ainsi que des relations d'auteur entre auteurs et documents.

La tâche de RI au sein d'un réseau bibliographique se décline par la recherche de documents pertinents ou auteurs autoritaires en considérant un sujet (requête). Deux grandes approches d'ordonnancement d'entités sont proposées dans la littérature : les approches bibliométriques et les approches basées sur la structure du réseau. Les approches bibliométriques ordonnancent les entités en se basant sur des indicateurs de qualité d'une publication (Walker *et al.*, 2006) ou d'importance d'un auteur (Egghe, 2006, Hirsch, 2005). Les approches basées sur la structure du réseau distinguent les modèles d'ordonnancement d'un seul type d'entités (Jabeur *et al.*, 2010, Kirsch *et al.*, 2006) ou de ceux qui ordonnent plusieurs types d'entités simultanément en tenant compte de la structure des relations dans le réseau (Yan *et al.*, 2010, Yang *et al.*, 2010, Zhang *et al.*, 2008, Zhou *et al.*, 2007).

Dans ce papier, nous proposons un modèle d'ordonnancement d'entités hétérogènes, associé à une fonction de pertinence, dans le cadre applicatif de la RI dans les

1. http ://twitter.com/
2. http ://www.flickr.com/
3. http ://www.linkedin.com/

réseaux bibliographiques. Ce modèle, appelé *BibRank*, permet d'estimer conjointement un score de pertinence pour chaque entité en considérant la structure du réseau. Chaque relation est pondérée, entre autres, par un indicateur de proximité thématique. Pour cela, nous appliquons le modèle de langue afin de relier deux sources d'évidence suivant le sens de la relation : l'expertise d'un auteur et le sujet d'un document.

Cet article est organisé comme suit. La section 2 introduit les approches de l'état de l'art connexes à l'ordonnancement d'entités bibliographiques. La section 3 définit les notions préliminaires ainsi que les notations. La section 4 présente le modèle BibRank applicable à un réseau bibliographique composé d'auteurs et de documents ainsi que les différents paramètres du modèle. La section 5 décrit l'évaluation expérimentale menée sur la base bibliographique de référence CiteSeerX. Enfin, nous terminons, en section 6, par une conclusion et annonçons nos perspectives de recherche.

2. Modèles d'accès à l'information dans les réseaux bibliographiques

Le domaine de la RI propose de nombreuses tâches de recherche, dont celle de l'estimation de la pertinence d'information vis-à-vis d'une requête, appelée couramment la recherche d'information *ad-hoc*. Nous déclinons par la suite les deux grandes approches de recherche *ad-hoc* citées précédemment dans les réseaux d'informations : l'approche bibliométrique et l'approche basée sur la structure du réseau.

L'approche bibliométrique ordonnance les entités bibliographiques en fonction de la qualité des publications scientifiques ou de l'importance des auteurs. Les indicateurs bibliométriques permettent ainsi de synthétiser les informations relatives à la production scientifique (Ibañez *et al.*, 2011). Ces mesures analysent les liens de citation entre entités (Alonso *et al.*, 2010, Zhang, 2009, Egghe, 2006, Hirsch, 2005) ou intègrent un aspect temporel supplémentaire lié à la date de publication (Bergstrom *et al.*, 2008, Garfield, 2006, Walker *et al.*, 2006). Par exemple, Hirsch (Hirsch, 2005) propose d'estimer l'importance d'un auteur en considérant l'ensemble de ses documents publiés et le nombre de citations de ces derniers. Walker et al. (Walker *et al.*, 2006) évaluent le nombre de citations qu'une publication scientifique peut recevoir à l'avenir compte tenu du nombre de citations reçues depuis sa date de publication.

L'approche basée sur la structure du réseau distingue deux principales catégories de modèles consistant, pour la première, à ordonnancer un seul type d'entités en tenant compte de leurs interactions avec l'ensemble des entités (Jabeur *et al.*, 2010, Kirsch *et al.*, 2006, Liu *et al.*, 2005) ou, pour la seconde, proposant un ordonnancement d'entités hétérogènes de façon conjointe considérant la structure du réseau (Yan *et al.*, 2010, Yang *et al.*, 2010, Tang *et al.*, 2008, Zhang *et al.*, 2008, Zhou *et al.*, 2007).

Parmi les travaux de la première catégorie, nous présentons trois modèles ordonnançant soit les documents (Jabeur *et al.*, 2010, Kirsch *et al.*, 2006), soit les auteurs (Liu *et al.*, 2005). Les modèles introduits par (Jabeur *et al.*, 2010, Kirsch *et al.*, 2006), considèrent le réseau comme un réseau social en ordonnançant les documents en fonction de leur pertinence thématique en réponse à un besoin en information et de la po-

sition sociale de leurs auteurs traduisant ainsi leur importance dans le réseau. Liu et al. (Liu *et al.*, 2005) présentent une extension de l'algorithme *PageRank* appliqué à un réseau de co-auteurs, où les relations sont pondérées par des fréquences de co-auteur, afin d'ordonner les auteurs en fonction de leur impact sur la production scientifique.

Les modèles de la deuxième catégorie utilisent une approche incluant une analyse des liens (Yan *et al.*, 2010, Zhou *et al.*, 2007) ou associant une dimension thématique supplémentaire (Yang *et al.*, 2010, Tang *et al.*, 2008, Zhang *et al.*, 2008).

Parmi les approches intégrant l'analyse des liens, Yan et Din (Yan *et al.*, 2010) proposent un modèle, appelé *PRank*, qui ordonnance conjointement les documents, auteurs et revues en se basant sur les hypothèses suivantes : (1) un important journal ainsi qu'un important auteur publient d'importants documents et (2) sa réciproque ; considérant cela, ce modèle propage le score des entités connectées hormis pour les documents dont le calcul du score diffère. En effet, le modèle *PRank* calcule un score de *PageRank* afin d'évaluer leur importance dans leur réseau homogène en considérant comme score initial le score généré à partir des hypothèses (1) et (2). Zhou et al. (Zhou *et al.*, 2007) proposent d'ordonnancer conjointement les entités du réseau en combinant un score de *PageRank*, reflétant ainsi l'autorité de l'entité dans son graphe homogène, et un score issu d'une fonction *BiWalk*, modélisant l'autorité de l'entité au travers des relations inter-graphes.

D'autres travaux (Yang *et al.*, 2010, Tang *et al.*, 2008, Zhang *et al.*, 2008) intègrent dans leur fonction d'ordonnancement un aspect thématique supplémentaire. Zhang et al. (Zhang *et al.*, 2008) proposent de recommander les entités en combinant un score thématique qui, grâce au modèle de langue, estime la similarité entre la requête et l'entité, et un score d'autorité de l'entité par rapport à l'ensemble du réseau hétérogène grâce à une extension de l'algorithme *PageRank*. Yang et al. (Yang *et al.*, 2010) proposent d'étendre le modèle *Topical PageRank*, introduit dans (Nie *et al.*, 2006), afin de mettre en valeur les entités autoritaires représentées par une distribution vectorielle de thèmes. Cet algorithme considère trois comportements pour un utilisateur : *"Follow-Stay"* lorsqu'un utilisateur navigue dans le réseau en restant dans le même thème, *"Follow-Jump"* lorsqu'il change de thème et *"Jump-Jump"* lorsqu'il accède de façon aléatoire à un thème. Le modèle *"Author-Topic-Conference"* (ACT), proposé dans (Tang *et al.*, 2008), est utilisé par le moteur de recherche Arnetminer[4]. Ce modèle représente les entités grâce à une distribution thématique inférée du modèle LDA, *"Latent Dirichlet Allocation"* (Wei *et al.*, 2006), et ordonnance ensuite les entités document, auteur et conférence en appliquant un algorithme proche de *PageRank*.

Comparativement aux travaux précédents, notre modèle en diffère en trois points :

– Nous proposons une approche intégrée afin d'ordonnancer de façon conjointe l'ensemble des entités du réseau et d'intégrer l'ensemble des facteurs du modèle dans un processus unifié, contrairement aux travaux (Jabeur *et al.*, 2010, Zhang *et al.*, 2008, Zhou *et al.*, 2007, Kirsch *et al.*, 2006) qui proposent une approche modulaire.

4. http ://www.arnetminer.com/

– De plus, notre approche intègre la structure du réseau ainsi que des indicateurs de proximité thématique contrairement aux travaux (Yan *et al.*, 2010, Yang *et al.*, 2010, Zhou *et al.*, 2007) qui considèrent seulement une analyse des liens entre entités. Dans ce travail, nous proposons, en effet, deux indicateurs de proximité thématique qui estiment la significativité d'une relation entre deux entités. Le premier indicateur applique le modèle de langue au contexte auteur-document afin de relier deux sources d'évidence reflétant le sujet du document et l'expertise de l'auteur. Le deuxième indicateur analyse l'intérêt commun des deux entités homogènes pour la requête et permet ainsi de discréditer les relations non significatives thématiquement, appelées aussi citations marginales. Cette approche s'oppose à celle de (Tang *et al.*, 2008) qui modélise les entités par des distributions thématiques.

– Enfin, l'idée principale de notre modèle demeure dans l'utilisation du modèle de langue afin d'estimer la significativité d'une relation entre entités, en plus de l'intégration de la structure du réseau comparativement aux travaux de (Zhang *et al.*, 2008) qui analyse de façon structurelle les relations entre entités et introduit la pertinence thématique seulement au niveau de la requête.

3. Notions préliminaires et notations

Le cadre applicatif de notre modèle est le suivant : ordonnancer conjointement les auteurs et les documents d'un réseau d'information bibliographique selon leur pertinence vis-à-vis d'une requête traduisant un sujet scientifique. À la différence des modèles d'appariement *ad-hoc* estimant la pertinence d'un document d_i par la fonction *Relevance Status Value* $RSV(Q, d_i)$, notre modèle ordonne conjointement les auteurs et les documents par une fonction de pertinence, appelée *BibRank*, tenant compte à la fois du sujet de la requête mais aussi de la structure du réseau. Ci-dessous, nous introduisons quelques définitions et notations.

– Le document : Un document scientifique, noté d_i est représenté par un vecteur de termes pondérés $\vec{d_i} = (w_{i1}, \ldots, w_{ip}, \ldots, w_{iT})$. w_{ip} représente le poids du $p^{ième}$ terme apparaissant dans le document d_i. Le nombre de termes de l'index est noté T.

– L'auteur : Un auteur scientifique a_j publie un ou plusieurs documents. Soit $\mathcal{D}(a_j)$ l'ensemble des documents publiés par a_j. Un auteur a_j est représenté par un vecteur de termes pondérés incluant les termes présents dans l'ensemble de ses documents publiés, $\vec{a_j} = \sum_{d_i \in \mathcal{D}(a_j)} \vec{d_i}$.

– Les graphes homogènes des documents et des auteurs : Les graphes homogènes des documents et des auteurs sont des graphes orientés incluant un ensemble de nœuds homogènes reliés par un seul type de relation. Le graphe homogène des documents, noté $G_D = (D, D \times D)$, représente l'ensemble des documents et de leurs relations possibles entre eux, appelées *relations intra-graphes*. De façon similaire, le graphe homogène des auteurs, noté $G_A = (A, A \times A)$, représente l'ensemble des auteurs des documents scientifiques et de leurs relations, dites également *relations intra-graphe*. Un relation (x, y) est appelée *intra-graphe* si les entités source $x \in X$ et cible

$y \in X$ sont incluent dans le même graphe homogène G_X avec $X \in \{A, D\}$. Considérant les graphes homogènes des documents et des auteurs, nous pouvons associer respectivement deux sémantiques à ces relations : la relation de citation $(d_i, d_k) \in D \times D$ entre documents d_i et d_k si le document d_i *cite* d_k et la relation $(a_j, a_l) \in A \times A$ de citation entre deux auteurs a_j et a_l si l'auteur a_j *cite* a_l par l'intermédiaire de ses documents publiés $\mathcal{D}(a_j)$.

– Le réseau bibliographique hétérogène : L'ensemble des ressources bibliographiques ainsi que leurs relations est modélisé par un graphe orienté $G = \{V, E\}$ où les nœuds $V = A \cup D$ représentent l'ensemble des entités bibliographiques avec D et A correspondant respectivement à l'ensemble des documents $D = \{d_1, \ldots, d_n\}$ et à l'ensemble des auteurs $A = \{a_1, \ldots, a_m\}$. L'ensemble des arcs $E = V \times V$ représente les relations bibliographiques entre les entités. La relation de l'entité $x \in V$ vers l'entité $y \in V$ est représentée par un arc orienté (x, y).

La figure 1 illuste un exemple d'un réseau bibliographique composé des deux sous-graphes homogènes des documents $G_D \in G$ et auteurs $G_A \in G$. Les relations permettant de connecter les deux sous-graphes homogènes sont appelées *relations inter-graphes*. Un relation (x, y) est appelée *inter-graphe* si les entités source $x \in X$ et cible $y \in Y$ appartiennent respectivement à deux graphes homogènes distincts G_X et G_Y avec $X \in \{A, D\}$, $Y \in \{A, D\}$ et $X \neq Y$. Dans notre réseau bibliographique, *les relations d'auteur* sont des relations inter-graphes où un auteur a_j et son document d_i sont connectés par une relation bi-directionnelle, modélisée selon deux arcs notés $(d_i, a_j) \in D \times A$ et $(a_j, d_i) \in A \times D$.

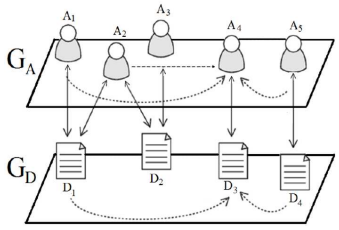


Figure 1 – Réseau bibliographique G

– La fonction *BibRank* : Soit un réseau bibliographique $G = (V, E)$ et une requête Q, la fonction *BibRank* ordonne conjointement les auteurs et les documents afin de produire une liste ordonnée pour chaque type d'entités.

$$BibRank : \{Q, G\} \longrightarrow \{R_A, R_D\} \quad [1]$$

$$\forall a_j \in A,\ 0 < R_A(a_j) < 1, \sum_{a_j \in A} R_A(a_j) = 1$$

$$\forall d_i \in D,\ 0 < R_D(d_i) < 1, \sum_{d_i \in D} R_D(d_i) = 1$$

avec R_D et R_A étant l'ensemble des scores des documents et des auteurs du réseau. $R_D(d_i)$ associe à chaque document d_i son score correspondant dans R_D. Respectivement, $R_A(a_j)$ associe à chaque auteur a_j son score correspondant dans R_A.

4. Ordonnancement conjoint de documents et d'auteurs dans un réseau bibliographique

Nous proposons le modèle *BibRank* ordonnant conjointement les auteurs et les documents d'un réseau bibliographique. Ce modèle est basé sur l'hypothèse de renforcement mutuel présenté dans de précédents travaux (Yang *et al.*, 2010, Zhang *et al.*, 2008), où un document, ou un auteur, important est cité par des documents, respectivement par des auteurs, importants. Ce renforcement mutuel est modélisé par un algorithme de propagation des scores entre entités connectées. Notre modèle intègre plusieurs facteurs permettant de considérer la structure du réseau à différents niveaux :

– L'analyse des relations entre sous-graphes des auteurs et documents est estimée au niveau global par la probabilité d'accéder à un sous-graphe homogène considérant la position d'un utilisateur. Cette probabilité est appelée *probabilité de transition*.

– L'analyse thématique des relations entre sous-graphes permet d'estimer au niveau granulaire la significativité des liens selon la thématique des documents, ou des auteurs, connectés. En effet, ce facteur permet d'analyser le lien entre deux documents, ou auteurs, plutôt que l'ensemble des relations du même type, comme pour les probabilités de transition. Cet indicateur est appelé *indicateur de proximité thématique*.

4.1. *Probabilités de transition*

Comme dans le cas de l'algorithme *PageRank*, nous modélisons grâce aux probabilités de transition le déplacement de l'utilisateur dans le graphe bibliographique. Lorsqu'un utilisateur accède à un nœud document, par exemple, nous énumérons deux actions possibles : rester dans le même sous-graphe pour accéder à un autre nœud document ou bien changer de sous-graphe afin d'accéder à un nœud auteur. Pour cela, nous estimons la probabilité de naviguer d'un sous-graphe G_X avec $X \in \{A, D\}$ à un autre sous-graphe G_Y avec $Y \in \{A, D\}$ par la probabilité de transition suivante :

$$\lambda_{XY} = \frac{|\{\forall (x, y) \in X \times Y\}|}{|E|} \; avec \; \lambda_{AD} + \lambda_{AA} = 1, \lambda_{DA} + \lambda_{DD} = 1 \quad [2]$$

où $|E|$ dénote le nombre de relations dans le réseau bibliographique. Au sein du graphe G, quatre probabilités de transition distinctes sont possibles : les probabilités intra-graphes (λ_{AA} et λ_{DD}) ainsi que les probabilités inter-graphes (λ_{AD} et λ_{DA}).

4.2. *Proximité thématique*

La proximité thématique est un indicateur basé sur le contenu permettant d'estimer la significativité d'une relation entre deux documents et/ou auteurs. Pour les relations d'auteur, le score de proximité thématique entre un document et son auteur est mesuré par un modèle de langue appliqué au contexte auteur-document qui relie, suivant le sens de la relation, le sujet d'un document et l'expertise d'un auteur. Pour les relations

de citation, la proximité thématique estime la significativité du lien de citation entre deux documents ou deux auteurs. Cet indicateur valorise ainsi les liens de citation significatifs et discrédite les citations marginales. Ce score est mesuré par l'intérêt commun des deux documents ou auteurs connectés par rapport au sujet de la requête.

4.2.1. *La représentativité du document par rapport à l'expertise de l'auteur*

Pour chaque auteur a_j, un indicateur de proximité thématique est attribué à chacun de ses documents publiés $\mathcal{D}(a_j)$. Ce score estime la vraisemblance pour un document d_i d'être en correspondance avec l'expertise de son auteur. Le score $ProxSem(d_i|a_j)$ d'un document d_i pour un auteur a_j est défini ainsi :

$$ProxSem(d_i|a_j) = \frac{P(a_j|M_{d_i})}{\max\limits_{\forall(a_l,d_k)\in A\times D} P(a_l|M_{d_k})} \quad [3]$$

avec $P(a_j|M_{d_i})$ étant la probabilité de pertinence d'un document d_i sachant un auteur a_j. Cette probabilité est définie selon un modèle de langue de la façon suivante :

$$P(a_j|M_{d_i}) = \prod_{t\in\vec{a_j}} [(1-\lambda)P(t|M_{a_j}) + \lambda P(t|M_{d_i})]^{n(t,\vec{a_j})} \quad [4]$$

$n(t,\vec{a_j})$ est le nombre d'occurrences du terme t dans $\vec{a_j}$. $P(t|M_{a_j})$ est la probabilité d'observer le terme t considérant le modèle de langue M_{a_j} de l'auteur a_j. $P(t|M_{d_i})$ est la probabilité d'observer le terme t considérant le modèle de langue M_{d_i} du document d_i. De façon générale, la probabilité $P(t|M_x)$ d'un terme t considérant le modèle de langue de l'entité $x \in A \cup D$ est défini comme suit : $P(t|M_x) = \frac{tf(t,x)}{|x|}$ où $tf(t,x)$ est la fréquence du terme t dans x et $|x|$, le nombre de termes inclus dans x.

4.2.2. *L'expertise de l'auteur par rapport au sujet du document*

Pour un document d_i, l'ensemble de ses auteurs est noté $\mathcal{A}(d_i)$. L'expertise de chaque auteur $a_j \in \mathcal{A}(d_i)$ considérant le sujet du document d_i est évaluée par l'indicateur $ProxSem(a_j|d_i)$:

$$ProxSem(a_j|d_i) = \frac{P(d_i|M_{a_j})}{\max\limits_{\forall(a_l,d_k)\in A\times D} P(d_k|M_{a_l})} \quad [5]$$

$P(d_i|M_{a_j})$ est la probabilité de pertinence d'un auteur a_j sachant un document d_i. Cette probabilité est définie selon un modèle de langue :

$$P(d_i|M_{a_j}) = \prod_{t\in\vec{d_i}} [(1-\lambda)P(t|M_c) + \lambda P(t|M_{a_j})]^{n(t,\vec{d_j})} \quad [6]$$

avec c étant la collection représentée par le vecteur de termes pondérés inclus dans tous les documents de la collection $c = \sum_{d_i\in D} \vec{d_i}$. $n(t,\vec{d_i})$ est le nombre d'occur-

rences du terme t dans $\vec{d_i}$. $P(t|M_c)$ est la probabilité d'observer le terme t considérant le modèle de langue M_c de la collection c. Cette probabilité est estimée ainsi : $P(t|M_c) = \frac{tf(t,c)}{|c|}$ où $tf(t,c)$ est la fréquence du terme t dans la collection c et $|c|$ représente le nombre de termes inclus dans la collection c.

4.2.3. *Citations marginales*

Les citations marginales sont détectées par l'étude de la significativité des liens de citation entre deux documents ou deux auteurs. Nous supposons que celle-ci peut être estimée grâce à la pertinence de chaque entité compte tenu du sujet de la requête. Pour cela, nous proposons de discréditer les citations marginales et mettre en valeur des liens de citation significatifs en comparant le rang $rang(x)$ et $rang(y)$ de chaque document ou auteur source x et cible y dans les résultats fournis par la fonction $RSV(Q,x)$ et respectivement $RSV(Q,y)$. Plus les rangs des deux documents ou des deux auteurs sont proches, plus les documents ou auteurs sont reliés par un lien de citation significatif. La similarité commune entre deux entités x et y d'un sous-graphe homogène G_X avec $X \in \{D, A\}$ est mesurée de la façon suivante :

$$Sim_{com}(x,y) = \frac{1}{|rang(x) - rang(y)|} \quad [7]$$

Parallèlement aux relations d'auteur, l'indicateur $ProxSem(y|x)$ d'une entité cible $y \in X$ à partir d'une entité source $x \in X$ avec $X \in \{D, A\}$ est estimée ainsi :

$$ProxSem(y|x) = \frac{Sim_{com}(x,y)}{\max\limits_{(v,w)\in X\times X} Sim_{com}(v,w)} \quad [8]$$

4.3. *Algorithme détaillé*

Le modèle *BibRank*, basé sur un algorithme de calcul de scores *PageRank*, ordonnance conjointement deux types d'entités : les documents et les auteurs. Cet algorithme intègre des indicateurs de proximité thématique ainsi que l'analyse des liens. Notre algorithme comporte trois principales étapes : 1) l'initialisation, 2) la propagation des scores selon les facteurs de pondération (probabilités de transition, proximité thématique et sujet de la requête) et 3) l'ordonnancement de chaque type d'entités.

Algorithm 1 BibRank

Données: $Q, G = (V, E), V = A \cup D$

Résultat: $BibRank : \{Q, G\} \longrightarrow \{R_D, R_A\}$

/* Étape 1: Initialisation */

$\theta \leftarrow 0$;

$R_D(d_i)^\theta \leftarrow \frac{1}{|D|}$;

$R_A(a_j)^\theta \leftarrow \frac{1}{|A|}$;

/* Étape 2: Propagation des scores en considérant les probabilités de transition, le sujet de la requête et la proximité thématique entre entités */

Répéter

/* Score des documents */

$$R_D(d_i)^{\theta+1} \leftarrow \frac{d}{|V|} + (1-d)\times \ (\lambda_{AD} \sum_{(a_l,d_i)\in A\times D} [\frac{R_A(a_l)^\theta * w_{a_l}^{d_i}}{O(a_l)}] + \lambda_{DD} \sum_{(d_k,d_i)\in D\times D} [\frac{R_D(d_k)^\theta * w_{d_k}^{d_i}}{O(d_k)}])$$

$R_D^{\theta+1} \leftarrow \mathcal{N}(R_D^{\theta+1})$;

/* Score des auteurs */

$$R_A(a_j)^{\theta+1} \leftarrow \frac{d}{|V|} + (1-d)\times \ (\lambda_{AA} \sum_{(a_l,a_j)\in A\times A} [\frac{R_A(a_l)^\theta * w_{a_l}^{a_j}}{O(a_l)}] + \lambda_{DA} \sum_{(d_k,a_j)\in D\times A} [\frac{R_D(d_k)^\theta * w_{d_k}^{a_j}}{O(d_k)}])$$;

$R_A^{\theta+1} \leftarrow \mathcal{N}(R_A^{\theta+1})$;

$\theta \leftarrow \theta + 1$;

jusqu'à *convergence*;

/* Étape 3: Ordonnancement des entités document et auteur */

$R_D^* \leftarrow Ord(R_D)$;

$R_A^* \leftarrow Ord(R_A)$;

avec :

– θ : le nombre d'itérations,

– $d \in [0, 1]$: le coefficient d'amortissement fixé à $d = 0.15$,

– $w_{a_l}^{d_i}$: $w_{d_k}^{d_i}$, $w_{a_l}^{a_j}$ et $w_{d_k}^{a_j}$: les facteurs de pondération pour les liens (a_l, d_i), (d_k, d_i), (a_l, a_j) et (d_k, a_j). De façon générale, les facteurs de pondération w_x^y d'une entité source $x \in A \cup D$ pour une entité cible $y \in A \cup D$ sont définis ainsi :

$$w_x^y = r_x * ProxSem(y|x) \quad [9]$$

r_x est le rang inverse de l'entité x, obtenu lors du calcul de la pertinence $RSV(Q, x)$,

– $O(d_k)$ et $O(a_l)$: le nombre de liens sortants du document d_k et de l'auteur a_l,

– $Ord(R_D)$ et $Ord(R_A)$: fonctions d'ordonnancement des documents et des auteurs en fonction des scores $R_D(d_i)$ et $R_A(a_j)$,

– $\mathcal{N}(R_D^\theta)$ et $\mathcal{N}(R_A^\theta)$: fonctions de normalisation des scores document et auteur.

La preuve de convergence de l'algorithme est liée de fait par celle de l'algorithme *PageRank* (Page *et al.*, 1999) puisqu'il est fondé sur une structure analogue. Le calcul du score de chaque entité, présenté dans l'algorithme 1, peut être exprimé matriciellement :

$$G_X = \frac{d}{|V|}e + (1-d)[\lambda_{XX} S_{XX} G_X + \lambda_{YX} S_{YX} G_Y] \quad [10]$$

G_X et G_Y représentent respectivement les vecteurs des scores des entités de type $X \in \{A, D\}$ et $Y \in \{A, D\}$ avec $X \neq Y$. e est un vecteur réel de taille $|X|$ où chaque composante est égale à 1. Les matrices $S_{XX} \in \mathbb{R}^{|X|\times|X|}$ et $S_{YX} \in \mathbb{R}^{|Y|\times|X|}$ correspondent respectivement aux matrices de transition des relations de citation et des relations d'auteur avec respectivement $S_{XX}(j,i) = \frac{w_{x_i}^{x_j}}{O(x_i)}$ ($j, k \in \{1, ..., |X|\}$) et $S_{YX}(j,k) = \frac{w_{y_k}^{x_j}}{O(x_k)}$ ($k \in \{1, ..., |Y|\}$).

5. Évaluation expérimentale

5.1. *Cadre expérimental*

5.1.1. *La collection bibliographique*

Nous avons utilisé la collection bibliographique pluridisciplinaire CiteSeerX[5] téléchargée via une application web en avril 2011. Le titre et le résumé de chaque document, la liste de ses auteurs ainsi que les liens de citations entre documents ont été retenus. Les liens de citation entre auteurs sont inférés de ces derniers en appliquant une correspondance exacte entre les noms des auteurs. Le Tableau 1 et la Figure 2 fournissent une analyse statistique de la collection et de son réseau bibliographique.

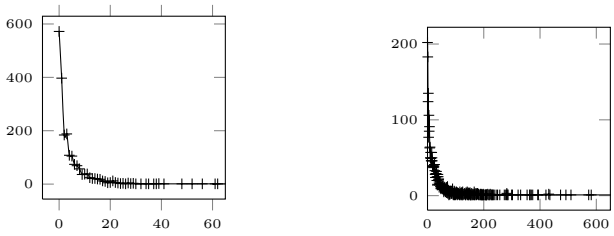
Documents	1 472 735
Auteurs	1 366 540
Liens de citation entre documents	16 598 502
Liens de citation entre auteurs	51 306 409
Liens de citations entre documents par document	11.270
Liens de citations entre auteurs par auteur	37.545
Liens d'auteur	4 209 980
Documents par auteur	3.081
Auteurs par document	2.858

Tableau 1 – Statistique sur la collection bibliographique CiteSeerX

5.1.2. *Requêtes*

L'ensemble de requêtes tests est construit à travers deux étapes :

5. http ://www.citeseer.ist.psu.edu/



(a) Réseau de citations entre documents

(b) Réseau de citations entre auteurs

Figure 2 – Analyse de la densité des relations de citations

– ***Étape 1*** *: Génération automatique des thèmes de la collection.* Pour cela, nous avons appliqué sur la collection l'algorithme LDA (Agichtein *et al.*, 2006) qui analyse la répartition thématique des termes et des documents afin de déterminer le nombre optimal de thèmes représentatifs de la collection. Cet algorithme fournit un ensemble de 200 thèmes représentés chacun par une liste de termes.

– ***Étape 2*** *: Dérivation manuelle des requêtes associées aux thèmes.* À partir des 200 thèmes extraits à l'Étape 1, nous avons identifié et éliminé les thèmes génériques. Chaque thème a été annoté manuellement, générant ainsi un sous-ensemble de 35 requêtes telles que les requêtes "web services ou "intelligent agent system". Pour chaque requête, nous avons extrait un sous-graphe incluant les premiers documents retournés par la fonction de pertinence RSV et leurs auteurs associés.

5.1.3. *Jugements de pertinence*

Pour chaque requête, nous avons besoin d'une valeur de pertinence pour l'ensemble des documents et des auteurs du sous-graphe extrait. Compte tenu que notre algorithme inclut à la fois le sujet de la requête et la structure du graphe, nous avons choisi de considérer deux indicateurs de pertinence binaires : un *indicateur thématique* et un *indicateur d'autorité*. En sommant ces deux indicateurs, nous obtenons ainsi un jugement de pertinence final à trois niveaux, compris entre 0 et 2.

– Jugement de pertinence des documents

L'indicateur thématique d'un document est obtenu par une expérimentation-utilisateur. Une liste des 20 premiers documents est d'abord générée pour chaque modèle d'ordonnancement. Les listes de résultats sont ensuite fusionnées. 25 participants de notre équipe, incluant assistants-professeurs, doctorants et étudiants en master, ont jugé la pertinence des documents par rapport au sujet de la requête. Chaque requête est évaluée par deux assesseurs différents qui attribuent un indicateur thématique binaire, lié au sujet de la requête. La mesure du Kappa de Cohen κ (Cohen, 1960) estime l'accord moyen entre les assesseurs pour l'ensemble des requêtes. Cette mesure est évaluée à une valeur de $\kappa = 0.57$ et reflète un accord modéré.

L'indicateur autorité d'un document est déduit d'une classification basée sur le critère du score de *PageRank*. La pertinence "autorité" d'un document est égale à 1 si le score

de *PageRank* du document est supérieur à la moyenne des scores de *PageRank* des documents dans leur sous-graphe homogène et 0 sinon.

– Jugement de pertinence des auteurs
La pertinence thématique d'un auteur est inférée de la moyenne de la pertinence thématique de ses documents.
L'indicateur autorité d'un auteur est déduit d'une classification basée sur le critère du score de *PageRank*. La pertinence "autorité" d'un auteur est égale à 1 si le score de *PageRank* de l'auteur est supérieur à la moyenne des scores de *PageRank* des auteurs dans leur sous-graphe homogène et 0 sinon.

5.1.4. *Modèles de référence*

Notre modèle a été comparé à trois modèles de l'état de l'art :

– **le modèle BM25** (Robertson *et al.,* 1994) représentant un modèle d'ordonnancement probabiliste, estimant la similarité entre une entité et le sujet de la requête.

– **le modèle de langue (LM) Hiemstra** (Hiemstra, 1998) représentant un modèle d'ordonnancement lissé. Pour rappel, nous avons utilisé le modèle de langue pour le facteur de proximité thématique, comme défini par les Formules 3 et 5.

– **le modèle PRank** (Yan *et al.,* 2010) représentant un modèle d'ordonnancement d'entités hétérogènes (documents, auteurs et journaux) basé sur l'analyse des liens. Ce modèle a été adapté à notre réseau social composé des nœuds document et auteur.

5.2. *Résultats et discussion*

Nous avons analysé l'efficacité de notre modèle pour l'ordonnancement des documents et des auteurs par rapport aux modèles de référence cités précédemment. Compte tenu que le jugement de pertinence d'une entité est évalué à trois niveaux, nous avons utilisé la mesure de $NDCG@20$, *Normalized Discount Cumulative Gain* (NDCG) (Agichtein *et al.,* 2006), qui intègre la mesure de pertinence graduelle afin de valoriser les systèmes d'ordonnancement proposant au début de la liste de résultats les entités les plus pertinentes. La significativité des accroissements de notre modèle a été évaluée en appliquant le test de Student. Le tableau 2 illustre les résultats obtenus par notre modèle pour les deux ordonnancements des documents et des auteurs.

Notre modèle *BibRank* permet d'atteindre un accroissement significatif pour les deux types d'ordonnancement des auteurs et des documents comparativement aux trois modèles de référence. Ces résultats montrent que la prise en compte séparée des facteurs "sujet de la requête", intégré par les modèles BM25 et LM, ou "structure du réseau", considéré par le modèle PRank, ne suffit pas pour une tâche d'ordonnancement d'entités hétérogènes dans un réseau bibliographique. En effet, notre modèle prend en compte simultanément le sujet de la requête ainsi que la structure du réseau, mais ajoute également un indicateur de proximité thématique entre entités connectées. L'introduction de l'ensemble de ces paramètres a permis de constater l'accroissement

Modèle	NDCG@20	Tx accroissemnt	
BM25	0.429	**59.77%**	***
LM	0.322	**113.13%**	***
PRank	0.641	**7.03%**	*
BibRank	**0.686**		

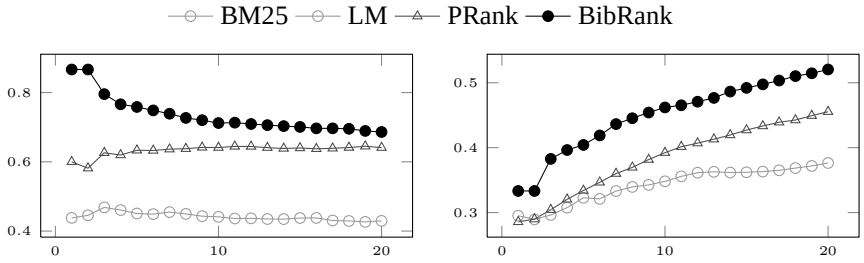
(a) Ordonnancement des documents

Modèle	NDCG@20	Tx accroissemnt	
BM25	0.376	**38.26%**	***
LM	0.428	**21.47%**	**
PRank	0.455	**14.29%**	*
BibRank	**0.520**		

(b) Ordonnancement des auteurs

Tableau 2 – Efficacité d'ordonnancement. % : accroissement de BibRank. Significativité du test de Student * : $0.01 < t \leqslant 0.05$; ** : $0.001 < t \leqslant 0.01$; *** : $t \leqslant 0.001$

significatif de notre modèle *BibRank* variant entre 7.03% et 113.13% pour l'ordonnancement des documents avec une mesure du NDCG de 0.686 et entre 14.29% et 38.26% pour celui des auteurs avec une valeur du NDCG à 0.520. Cependant, l'accroissement est plus faible pour le modèle *PRank*. En effet, ce modèle répond à un même besoin que notre modèle *BibRank* : ordonnancer des entités hétérogènes d'un réseau bibliographique. Nous remarquons, de plus, que les valeurs du $NDCG@20$ sont généralement plus faibles pour l'ordonnancement des auteurs que pour celui des documents. Une explication possible peut résider dans la façon de déterminer les jugements de pertinence thématique pour les auteurs. En effet, ces derniers sont déduits automatiquement de ceux des documents et ne sont pas le résultat d'un jugement humain. La figure 3 illustre les courbes de NDCG pour l'ordonnancement des deux entités à différents niveaux de rang, compris entre 1 et 20. Nous observons que les deux courbes de notre modèle *BibRank* sont supérieures à celles des modèles de référence. L'ordonnancement est particulièrement efficace pour les documents quand on considère les 5 premiers ; ce qui implique que notre modèle place les documents les plus pertinents au début de la liste de résultats.



(a) Ordonnancement des documents

(b) Ordonnancement des auteurs

Figure 3 – La mesure du NDCG à différents niveaux d'ordonnancement

6. Conclusion

Dans ce papier, nous proposons un modèle d'ordonnancement d'entités dans un réseau bibliographique, appelé *BibRank*. Ce modèle propage le score des documents et des auteurs en considérant à la fois la structure du réseau et la thématique des entités

bibliographiques. Pour cela, nous estimons la significativité thématique des liens de citation, l'expertise d'un auteur par rapport au sujet d'un document ou encore la représentativité d'un document par rapport à l'expertise de son auteur. Nous avons mené une série d'expérimentations sur la collection CiteSeerX qui a montré un accroissement significatif du modèle *BibRank* par rapport aux autres modèles de l'état de l'art pour l'ordonnancement des deux types d'entités. Nos futurs travaux sont orientés à court terme sur l'extension notre modèle à un réseau bibliographique incluant un plus grand nombre de types d'entités, tels que les conférences ou les institutions scientifiques, et de relations, comme les relations de co-auteur ou d'affiliation. À moyen terme, nous souhaitons appliquer notre modèle à un réseau social du web afin d'intégrer des facteurs temporels ou sociaux. L'importance d'une entité ne dépendrait plus de son importance en terme de liens de citations, mais plutôt de son autorité moyenne depuis sa date de publication ou de la distance sociale la séparant des autres entités.

7. Bibliographie

Agarwal N., Liu H., Tang L., Yu P. S., « Identifying the influential bloggers in a community », *Proceedings of the international conference on Web search and web data mining*, WSDM '08, ACM, p. 207-218, 2008.

Agichtein E., Brill E., Dumais S., « Improving web search ranking by incorporating user behavior information », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 19-26, 2006.

Alonso S., Cabrerizo F. J., Viedma E. H., Herrera F., « hg-index : a new index to characterize the scientific output of researchers based on the h- and g-indices. », *Scientometrics*, vol. 82, p. 391-400, 2010.

Bergstrom C. T., West J. D., Wiseman M. A., « The Eigenfactor metrics. », *Journal of Neuroscience*, vol. 28, n° 45, p. 11433-11434, 2008.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, n° 1, p. 37-46, 1960.

Egghe L., « An improvement of the h-index : the g-index », *ISSI Newsletter*, vol. 2, p. 8-9, 2006.

Emmerink R., Effects of information in road transport networks with recurrent congestion, Technical report, 1993.

Garfield E., « The History and Meaning of the Journal Impact Factor », *JAMA : The Journal of the American Medical Association*, 2006.

Hiemstra D., « A Linguistically Motivated Probabilistic Model of Information Retrieval », *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, Springer-Verlag, London, UK, p. 569-584, 1998.

Hirsch J. E., « An index to quantify an individual's scientific research output », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, p. 16569-16572, 2005.

Ibañez A., Larrañaga P., Bielza C., « Using Bayesian networks to discover relationships between bibliometric indices. A case study of computer science and artificial intelligence journals », *Scientometrics*, 2011.

Jabeur L. B., Tamine L., Boughanem M., « A social model for literature access : towards a weighted social network of authors », *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, p. 32-39, 2010.

Kirsch S., Gnasa M., Cremers A., « Beyond the Web : Retrieval in Social Information Spaces », *Advances in Information Retrieval*p. 84-95, 2006.

Liu X., Bollen J., Nelson M. L., Van de Sompel H., « Co-authorship networks in the digital library research community », *Inf. Process. Manage.*, vol. 41, p. 1462-1480, 2005.

Nie L., Davison B. D., Qi X., « Topical link analysis for web search », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 91-98, 2006.

Page L., Brin S., Motwani R., Winograd T., The PageRank Citation Ranking : Bringing Order to the Web., Technical Report n° 1999-66, Stanford InfoLab, November, 1999.

Robertson S. E., Walker S., « Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval », *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, Springer-Verlag New York, Inc., New York, USA, p. 232-241, 1994.

Rorissa A., « A comparative study of Flickr tags and index terms in a general image collection », *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, p. 2230-2242, November, 2010.

Roy S., Lane T., Werner-Washburne M., « Integrative construction and analysis of condition-specific biological networks », *Proceedings of the 23rd national conference on Artificial intelligence*, vol. 3, AAAI Press, p. 1867-1868, 2008.

Tang J., Jin R., Zhang J., « A Topic Modeling Approach and Its Integration into the Random Walk Framework for Academic Search », *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, Washington, DC, USA, p. 1055-1060, 2008.

Walker D., Xie H., Yan K.-K., Maslov S., « Ranking Scientific Publications Using a Simple Model of Network Traffic », *Society*p. 1-5, 2006.

Wei X., Croft W. B., « LDA-based document models for ad-hoc retrieval », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, ACM, New York, NY, USA, p. 178-185, 2006.

Yan E., Ding Y., « Measuring scholarly impact in heterogeneous networks », *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, American Society for Information Science, Silver Springs, MD, USA, p. 1-7, 2010.

Yang Z., Hong L., Davison B. D., « Topic-driven multi-type citation network analysis », *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, Paris, France, p. 24-31, 2010.

Zhang C.-T., « The e-Index, Complementing the h-Index for Excess Citations », *PLoS ONE*, 2009.

Zhang J., Tang J., Liang B., Yang Z., Wang S., Zuo J., Li J., « Recommendation over a Heterogeneous Social Network », *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, WAIM '08, IEEE Computer Society, Washington, DC, USA, p. 309-316, 2008.

Zhou D., Orshanskiy S. A., Zha H., Giles C. L., « Co-ranking Authors and Documents in a Heterogeneous Network », *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC, USA, p. 739-744, 2007.