Modèle de Recherche d'Information Sociale Centré Utilisateur

Chahrazed Bouhini, Mathias Géry, Christine Largeron

Modèle de Recherche d'Information Sociale Centré Utilisateur

Chahrazed Bouhini*, Mathias Géry*
Christine Largeron*

*Université de Lyon, F-42023, Saint-Étienne, France;
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42023, Saint-Étienne, France;
Université de Saint-Étienne, Jean-Monnet, F-42023, Saint-Étienne, France.
{Chahrazed.Bouhini, Mathias.Gery, Christine.Largeron}@univ-st-etienne.fr
http://www.laboratoirehubertcurien.fr

Résumé. L'émergence des réseaux sociaux a révolutionné le Web en permettant notamment aux individus de prolonger leur connexion virtuelle en une relation plus réelle et de partager leurs connaissances. Ce nouveau contexte de diffusion de l'information sur le Web peut constituer un moyen efficace pour cerner les besoins en information des utilisateurs du Web, et permettre à la recherche d'information (RI) de mieux répondre à ces besoins en adaptant les modèles d'indexation et d'interrogation. L'exploitation des réseaux sociaux confronte la RI à plusieurs défis dont les plus importants concernent la représentation de l'information dans ce modèle social de RI et son évaluation, en l'absence de collections de test et de compétitions dédiées. Dans cet article, nous présentons un modèle de RI sociale dans lequel nous proposons de modéliser et d'exploiter le contexte social de l'utilisateur. Nous avons évalué notre modèle à l'aide d'une collection de test de RI sociale construite à partir des annotations du réseau social de bookmarking collaboratif *Delicious* [1].

1 Introduction

La recherche d'information (RI) est un domaine qui consiste à définir des modèles et des processus dont le but est de retourner, à partir d'un corpus de documents indexés, ceux dont le contenu correspond le mieux au besoin en information exprimé par un utilisateur. Initialement développée pour des corpus de documents textuels, la RI a évolué avec l'émergence du Web et plus récemment des réseaux sociaux (RS). De nos jours, les RS représentent le moyen le plus utilisé pour la communication, le partage de connaissance et de contenus sur le Web.

Avec cette dimension sociale qui vient enrichir les contenus des ressources sur le Web, les utilisateurs se retrouvent avec de nouveaux besoins en information. La RI classique ne semble pas adaptée à cette dimension, impliquant les utilisateurs et leurs interactions au sein des réseaux sociaux, d'où l'émergence de la RI Sociale (RIS), une thématique récente qui a pour objectif de prendre en compte les informations spécifiques aux RS.

1. Delicious : http://delicious.com

1.1 Les Réseaux Sociaux (RS)

Les réseaux sociaux sont un espace dans lequel les internautes interagissent (publient, partagent, annotent, commentent, etc.) avec le contenu du Web (Fischer et Reuber, 2010). Il peut s'agir d'images (Flickr : 6 miliards de photos [2]), de ressources (Twitter : plus de 500 millions d'utilisateurs, Facebook : plus d'un miliard d'utilisateurs [3], del.icio.us), ou encore d'informations professionnelles (LinkedIn : 175 millions de membres [4]). Les réseaux sociaux représentent aussi un moyen de communication et d'échange efficace en permettant aux utilisateurs de rentrer en contact avec des collègues, amis, co-auteurs et *followers*.

Dans une problématique d'accès à l'information, on peut distinguer trois principaux éléments constitutifs des RS, en plus des utilisateurs :

- Le contenu des documents (ressources) partagés, publiés ou annotés au sein du réseau ;
- Les annotations que les utilisateurs associent aux ressources ;
- Les relations sociales entre les utilisateurs du réseau reflétant la popularité (followers, fans, amis, etc.), la confiance (partage, types de commentaires, etc) ou l'expertise.

1.2 La Recherche d'Information Sociale (RIS)

La RIS a comme objectifs, dans un premier temps, d'améliorer le processus de RI en exploitant les informations sociales (cf. figure 1) et, dans un second temps, de personnaliser la recherche de l'utilisateur selon son contexte social. La figure 1 montre que les documents pertinents pour une requête donnée peuvent être différents d'un utilisateur à l'autre.

Les principales problématiques de cette thématique sont donc d'identifier, d'exploiter et de combiner les informations sociales issues des RS pour améliorer et éventuellement personnaliser la RI. Dans le travail présenté dans cet article, nous nous intéressons uniquement aux problématiques d'exploitation et de combinaison d'une partie des informations sociales (les annotations) dans l'objectif de construire le contexte social de l'utilisateur afin d'individualiser la RI en l'adaptant aux profils des utilisateurs. Ceci nous a conduit à proposer un modèle de RIS intégrant ce contexte social. Les annotations sociales fournissent à l'utilisateur un moyen d'enrichir à la fois son profil informationnel en annotant des ressources (expression d'une opinion, d'un certain niveau de connaissance sur un domaine, des centres d'intérêt, etc.) et le contenu informationnel des ressources annotées (par le biais d'une brève description personnelle, catégorisation personnelle de certaines ressources, etc.).

Dans la section suivante, nous allons commencer par positionner notre travail par rapport aux travaux d'état de l'art. Dans la section 3, nous présentons notre modèle de RIS centré utilisateur, puis les expérimentations réalisées et les résultats obtenus dans les sections 4 et 5.

2 État de l'Art

Dans les domaines connexes à la RI, de nombreux travaux ont abordé l'exploitation des réseaux sociaux dans les systèmes de recommandation (Guy et al. (2010), Konstas et al. (2009), etc.) et le filtrage collaboratif (Ferrara et Tasso, 2011).

2. http ://fr.slideshare.net/WaveLab/10-social-networks-and-a-bunch-of-stats
3. http ://www.zdnet.fr/actualites/facebook-un-milliard-d-utilisateurs-actifs-dans-le-monde-39783232.htm
4. http ://press.linkedin.com/about

Les recherches récentes en RI se sont focalisées sur l'utilisation des annotations sociales et/ou relations sociales pour l'amélioration des résultats de RI et le reclassement des documents par l'intégration de ces annotations dans les différentes méthodes habituelles de calcul de score (Wen et al. (2012), Vallet et al. (2010)).

Certains travaux proposent d'utiliser, pour l'expansion de la requête et l'amélioration des résultats de RI :

- Les annotations sociales : en tenant compte des co-occurrences des termes dans les annotations des utilisateurs (Theobald et al., 2005) et (Lin et al., 2011), en se basant sur le calcul de similarité des pages Web et les annotations sociales et la similarité sémantique de ces annotations (Bao et al., 2007).
- Les annotations et les relations sociales : en proposant d'exploiter la similarité des liens sociaux pour déterminer l'ensemble des annotations pouvant constituer un sous-ensemble de termes pour l'expansion de la requête (Schenkel et al., 2008).

D'autres travaux proposent de reclasser les ressources ou documents sur le Web en intégrant l'information sociale par combinaison des scores : un score social et un score thématique classique d'un document par rapport à une requête (Kirsch, 2005), où le score social peut être obtenu à partir des annotations et relations sociales (Ben Jabeur et al., 2010).

Mais, dans le cadre de la RI textuelle, Robertson et al. (2004) ont montré qu'il n'était ni cohérent d'un point de vue théorique, ni efficace d'un point de vue expérimental de réaliser une telle intégration directement par une combinaison de score. Robertson et al. (2004) proposent donc d'intervenir directement au niveau de la fonction de pondération (cf section 3.3).

Dans notre modèle de RIS, nous nous sommes intéressés aux méthodes et approches d'intégration de l'information sociale par combinaison de termes : les termes du contenu des ressources et les termes appartenant à l'information sociale. À notre connaissance, il n'existe pas dans la littérature de travaux dédiés à l'intégration de l'information sociale à l'étape d'indexation et de pondération des ressources.

3 Modèle de RI Sociale

En RIS, comme le montre la figure 1, les utilisateurs et les ressources sont caractérisés par des informations sociales, par exemple le contenu informationnel social des utilisateurs décrit par des annotations. Les informations sociales (IS) présentent une source d'informations supplémentaires décrivant à la fois les utilisateurs et les ressources. De ce fait, cette source d'information peut être exploitée pour décrire la requête ou le profil de l'utilisateur et/ou pour enrichir la description du contenu des ressources sur le Web. Dans notre cas, nous nous intéressons à l'exploitation de ces IS pour modéliser le contexte informationnel social de l'utilisateur et affiner la description des ressources du Web.

3.1 Notations

On modélise les données de la RIS par un quintuplet $< U, R, A, Q, Qrels >$:

- $U = \{u_1, u_2, ..., u_x, ..., u_{|U|}\}$ est l'ensemble des utilisateurs du réseau.
- $R = \{r_1, r_2, ..., r_i, ..., r_{|R|}\}$ représente une collection de ressources sur le Web (une image, une vidéo, une page Web, etc.). Au sein du réseau, les ressources R sont indexées par des termes d'index $t_j \in T$.

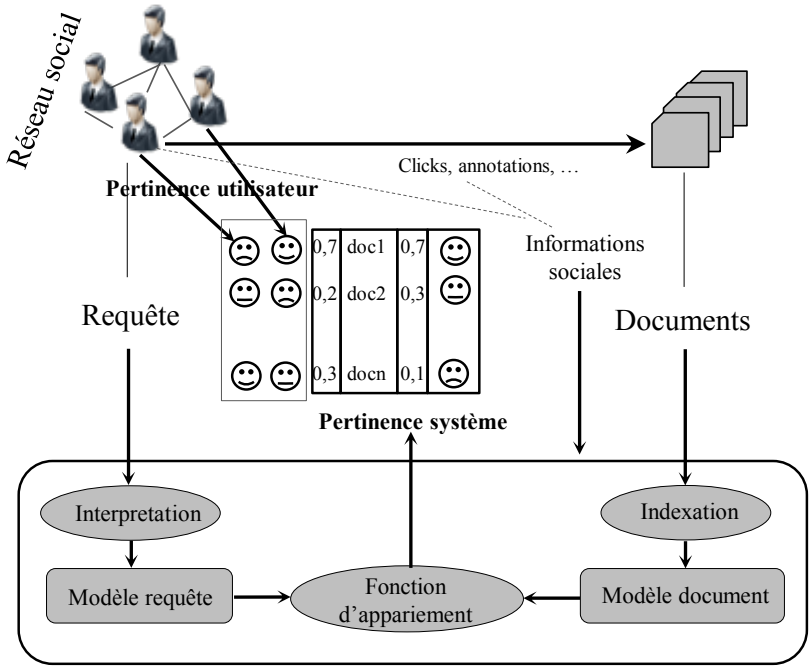

FIG. 1 – *Modèle de RIS.*

- Poids des termes $w_{i,j}$: à partir du nombre d'occurrences $tf_{i,j}$ de $t_j \in T$ dans la ressource $r_i \in R$, on calcule le poids $w_{i,j}$ de t_j pour r_i.
- $A = \{a_1, a_2, ..., a_z, ..., a_{|A|}\}$ représente l'ensemble des actions d'annotation des ressources R par les utilisateurs, avec $a_z =< T_z, r_i, u_x > \in A$ est l'annotation d'une ressource r_i par un utilisateur u_x avec des termes $T_z \subseteq T$.
- Q est l'ensemble des requêtes globales des utilisateurs, chaque requête q_l est composée d'un ou plusieurs termes $t_j \in T$.
- Q_{CU} est l'ensemble des requêtes centrée-utilisateur. Chaque requête $q_{l.u_x}$ correspond à une requête globale q_l posée par un utilisateur u_x.
- $Qrels$ est l'ensemble des jugements de pertinence : à chaque requête globale q_l est associée une liste de ressources pertinentes $qrels_l \subseteq R$.
- $Qrels_{l.u_x}$ est l'ensemble des jugements de pertinence centrée-utilisateur : à chaque requête centrée-utilisateur $q_{l.u_x}$ est associée une liste de ressources pertinentes $qrels_{l.u_x} \subseteq R$, qui peut être différente selon les utilisateurs.

3.2 Modélisation du contexte informationnel social

Le contexte informationnel social d'un utilisateur au sein des RS est composé de différents types d'informations (annotations, commentaires, traces, citations, voisinage réseaux, etc.). Dans le cadre de ce travail, le contexte informationnel social $Ctxt(u_x)$ d'un utilisateur est

obtenu à partir de ses annotations sociales uniquement. Il est représenté par l'ensemble des termes employés par u_x pour annoter des ressources sur le Web (cf. formule 1).

$$Ctxt(u_x) = \{t_j \in T_z, \forall a_z = < T_z, r_i, u_x > \in A(u_x)\} \quad (1)$$

avec : $A(u_x)$ désigne l'ensemble des annotations que l'utilisateur u_x a employé pour annoter les ressources.

Pour une ressource r_i donnée, le contexte informationnel social est donné par la formule suivante :

$$Ctxt(u_x, r_i) = r_i \cap Ctxt(u_x) \quad (2)$$

3.3 Indexation : intégration du contexte social

L'adaptation de la RI à chaque utilisateur nécessite d'indexer son contexte informationnel social. Nous proposons deux modèles d'indexation intégrant ce contexte ($BM25\_S$ et $BM25F\_S$), que nous comparons à un modèle d'indexation classique ($BM25$) qui n'intègre pas ce contexte.

Dans les modèles $BM25\_S$ et $BM25F\_S$, la ressource est composée de 2 champs : le champ contenu textuel de la ressource et le champ contexte de l'utilisateur adapté à cette ressource $Ctxt(u_x, r_i)$. Nous considérons ce dernier comme un champ apportant une description supplémentaire de la ressource, et nous proposons donc d'intégrer le contexte social de l'utilisateur au niveau de l'indexation des ressources. Une ressource est jugée pertinente par le système si le terme recherché se trouve dans le contexte de l'utilisateur et dans la ressource en question.

3.3.1 Modèle de référence : $BM25$

Le modèle d'indexation BM25 (Robertson et Walker, 1994) est un des modèles donnant les meilleurs résultats en RI documentaire classique dans les grandes compétitions de RI (INEX [5], TREC [6], etc.). Le poids d'un terme t_j dans une ressource r_i est calculé avec la formule 3 :

$$w_{i,j} = \frac{(k_1 + 1) \times tf_{i,j}}{k_1 \times ((b-1) + b \times (\frac{rl}{avgrl}) + tf_{i,j}} \times log(\frac{N - df_j + 0,5}{df_j + 0,5}) \quad (3)$$

avec :

- rl : la taille de la ressource r_i et $avgrl$: la taille moyenne des ressources.
- $tf_{i,j}$: le nombre d'occurrences du terme t_j dans la ressource r_i.
- k_1 : un paramètre qui permet de contrôler la saturation de $tf_{i,j}$.
- b : un paramètre qui permet de contrôler la normalisation par rapport à la taille des ressources.
- N : le nombre de ressources dans la collection.
- df_j : le nombre de ressources qui contiennent le terme t_j.

Dans ce modèle, le score global d'un document pour une requête est uniquement un score thématique classique. Le calcul de ce score est le suivant :

5. https ://inex.mmci.uni-saarland.de/
6. http ://trec.nist.gov/

$$Score_{BM25}(q_l, r_i) = \sum_{t_j \in r_i \cap q_l} w_{i,j} \quad (4)$$

3.3.2 BM25 avec contexte intégré : $BM25\_S$

Dans le modèle $BM25\_S$ que nous proposons, le score global d'un document prend en considération, en plus du score thématique classique, l'apport du contexte social.

Dans ce modèle $BM25\_S$, nous intégrons le contexte social par une simple concaténation des termes du contexte $Ctxt(u_x)$ avec les termes du contenu de la ressource. Nous utilisons ensuite le même modèle d'indexation $BM25$ que pour le modèle de référence, mais pour indexer cette fois-ci des ressources enrichies par le contexte social des utilisateurs. Cette indexation produit un index centré-utilisateur des ressources.

$$Score_{BM25\_S}(q_{l.u_x}, r'_i, u_x) = \sum_{t_j \in r'_i \cap q_l} w_{i,j} \quad (5)$$

Tel que r'_i est la ressource r_i enrichie par $Ctxt(u_x)$ adapté à cette ressource ($Ctxt(u_x, r_i)$) lors de l'indexation.

3.3.3 BM25 avec contexte intégré *a priori* : $BM25F\_S$

Afin de contrôler l'importance accordée au contexte par rapport au contenu, on peut envisager une combinaison *a posteriori* ou *a priori* des deux champs : contenu et contexte.

– *A posteriori*, c'est-à-dire une fois que les ressources ont été indexées, par combinaison du score thématique et du score social calculé à partir des informations sociales :

$$Score_{posteriori}(q_{l.u_x}, r_i, u_x) = \alpha Score_{BM25}(q_l, r_i) + (1-\alpha) Score_S(q_l, Ctxt(u_x)) \quad (6)$$

– *A priori*, c'est-à-dire avant l'indexation des ressources, par combinaison du nombre d'occurrences des termes apparaissant dans le contexte social avec le nombre d'occurrences des termes apparaissant dans le contenu de la ressource (cf. formule 7) :

$$ws_{x,i,j} = \frac{(k_1+1) \times (\alpha tf_{i,j} + (1-\alpha) tfs_{x,i,j})}{k_1 \times ((b-1) + b \times (\frac{rl}{avgrl}) + (\alpha tf_{i,j} + (1-\alpha) tfs_{x,i,j})} \times log(\frac{N - df_j + 0,5}{df_j + 0,5}) \quad (7)$$

avec $tfs_{x,i,j}$ la fréquence d'apparition du terme t_j dans le contexte ($Ctxt(u_x, r_i)$) et $ws_{x,i,j}$ le poids du terme t_j dans la ressource r_i pour l'utilisateur u_x.

Le score global d'une ressource pour une requête est calculé comme suit :

$$Score_{BM25F\_S}(q_{l.u_x}, r'_i, u_x) = \sum_{t_j \in r'_i \cap q_l} ws_{x,i,j} \quad (8)$$

Dans cet article, nous étudions uniquement la combinaison *a priori* des termes (score $BM25F\_S$). En effet, dans Robertson et al. (2004), les auteurs montrent dans le cas de documents composés de différents champs (résumé, corps du document, etc.), qu'une combinaison *a priori* (nommée $BM25F$ (F = Field)) est à la fois plus cohérente d'un point de vue théorique qu'une combinaison a posteriori, et plus efficace d'un point de vue expérimental. Ce principe

a été vérifié dans d'autres contextes, par exemple : l'intégration du poids des balises pour la RI (Géry et Largeron, 2012), et des tf des mots traduits dans différentes langues (Li et Gaussier, 2012). Dans notre modèle, nous avons choisi d'adapter la variante de $BM25F$ proposée dans Zaragoza et al. (2004), qui ajustent le paramètre b champ par champ, contrairement à Robertson et Walker (1994), qui ajustent un paramètre b global.

4 Expérimentations

4.1 Collection de test de RI Sociale

Les premières collections de test proposées pour évaluer la RIS (ex : TRECMicroblog [7], INEX Book Search [8], MendeleyDataTel (Jack et al., 2012)) prennent en compte les utilisateurs mais aucune d'entre elles ne contient de jugements de pertinence CU. Nous évaluons notre modèle de RIS à l'aide d'une collection de test dédiée, que nous avons construite, incluant des requêtes et des jugements de pertinence centrée-utilisateur et basée sur des données sociales du réseau *Delicious*. *Delicious* est un système collaboratif d'annotation qui permet aux utilisateurs d'annoter et catégoriser librement leur ressources (des pages Web) avec les mots clés appelés tags.

Le tableau 1 présente les caractéristiques de la collection $DelRIS$, composée de données classiques en RI (ressources, requêtes globales, jugements de pertinence globale) et de données spécifiques à la RIS (utilisateurs, requêtes CU, jugements de pertinence CU).

	Nombre	RI / RIS
Ressources	30224	RI
Utilisateurs	360	RIS
Requêtes globales	79	RI
Requêtes centrées-utilisateur	2049	RIS
Jugements de pertinence globale	4685	RI
Nombre moyen de ressources pertinentes globales	80	RI
Nombre moyen de ressources pertinentes CU	3	RIS

TAB. 1 – *Données de la collection de RIS* $DelRIS$.

- La collection contient 30224 ressources.
- 79 requêtes globales auxquelles correspondent 2049 requêtes centrées-utilisateur. Il y a en moyenne 25 requêtes centrées-utilisateur pour chaque requête globale.
- 4685 ressources pertinentes pour l'ensemble des 79 requêtes globales, donc en moyenne environ 80 ressources pertinentes par requête. Ces 79 requêtes globales se déclinent en 2049 requêtes CU (couples <utilisateur - requête>). Il y a donc en moyenne moins de 3 ressources pertinentes par requête CU.

Le faible nombre de ressources pertinentes par requête CU peut poser problème pour l'évaluation de la RIS. Pour palier ce problème, nous avons utilisé une variante de la collection

7. http ://trec.nist.gov/pubs/call2012.html
8. https ://inex.mmci.uni-saarland.de/tracks/books/

$DelRIS$, en conservant seulement les requêtes CU pour lesquelles il existe au minimum 10 ressources pertinentes. Le tableau 2 présente les caractéristiques de cette collection $DelRIS2$.

	Nombre	RI / RIS
Utilisateurs	70	RIS
Requêtes centrées-utilisateur	244	RIS
Jugements de pertinence globale	4685	RI
Nombre moyen de ressources pertinentes CU	10	RIS

TAB. 2 – *Données de la collection de RIS* $DelRIS2$.

Dans le cadre des expérimentations présentées dans cet article, nous avons utilisé seulement les requêtes correspondant à 2 utilisateurs, afin d'évaluer le modèle de référence ***BM25*** et nos deux modèles de RIS ($BM25\_S$ et $BM25F\_S$). Enfin, la collection comporte également 35698 annotations utilisant 21284 termes différents. L'annotation d'une ressource est un triplet <utilisateur, url, (tag_1, tag_2, ...)>. Les annotations sont utilisées pour construire le contexte informationnel social des utilisateurs (cf. section 3.2).

4.2 Mesures d'évaluation de RIS

En RI, les mesures d'évaluation classiques permettant d'évaluer la qualité des systèmes de RI sont le rappel, la précision et la MAP :

- Rappel : la capacité du système à retrouver toutes les ressources pertinentes (taux de ressources pertinentes retrouvées) : $Rappel = \frac{|PR|}{|P|}$
- Précision : la capacité à ne retrouver que les ressources pertinentes (taux des ressources pertinentes dans les ressources retrouvées) : $Precision = \frac{|PR|}{|PR \bigcup NPR|}$
  avec P : l'ensemble de ressources pertinentes, PR : l'ensemble de ressources pertinentes retrouvées et NPR : ensemble de ressources non pertinentes retrouvées.
- Précision moyenne AP_l : la moyenne des précisions à chaque ressource pertinente retournée pour une requête q_l (cf. equation 9).
- Moyenne des précisions moyennes MAP (***Mean Average Precision***) : la moyenne des AP_l pour toutes les requêtes permet d'obtenir une mesure de la performance globale du système (Manning et al., 2008).

$$AP_l = \frac{\sum_{k=1,K}(rel(k) * P@k)}{\sum_{k=1,K} rel(k)} \qquad MAP = \frac{1}{|Q|} \sum_{l=1,|Q|} AP_l \tag{9}$$

avec K le nombre de ressources retrouvées, $P@k$: la précision au rang k et $rel(k)$ calculé comme suit : $rel(k) = \begin{cases} 1 : si\ la\ k^{ieme}\ ressource\ est\ pertinente \\ 0 : sinon. \end{cases}$

Dans le cas de la RIS, nous avons proposé d'adapter la mesure MAP en calculant la moyenne des précisions moyennes $AP_{q_{l.u_x}}$ obtenues pour chaque couple <requête - utilisateur>. Cette mesure est appelée MAP_S. De la même manière, on définit $P_S[0.1]$, la précision

à un taux de rappel de 10%.

$$MAP_S = \frac{1}{|Q_{CU}|} \sum_{Q_{CU}} AP_{q_{l.u_x}} \qquad P_S[0.1] = \frac{1}{|Q_{CU}|} \sum_{Q_{CU}} P[0.1]_{q_{l.u_x}} \tag{10}$$

5 Résultats

Nous avons expérimenté le modèle de référence $BM25$ et nos deux modèles de RIS $BM25\_S$ et $BM25F\_S$. Les meilleurs résultats pour chaque mesure sont présentés en gras.

Le tableau 3 présente les résultats des 3 modèles sur la collection $DelRIS$, pour les mesures MAP_S et $P_S[0.1]$.

	MAP_S	$P_S[0.1]$
BM25	0,0511	0,0787
BM25_S	0,0529	**0,1413**
BM25F_S	**0,0546**	0,1126

TAB. 3 – *Évaluation avec la collection* $DelRIS$.

On constate une amélioration des résultats en terme de MAP_S et $P_S[0.1]$ quand le contexte est pris en compte. Cette amélioration est plus importante avec le modèle $BM25F\_S$ qu'avec le modèle $BM25\_S$ quand la MAP est considérée, mais c'est le contraire quand la mesure orientée précision $P_S[0.1]$ est considérée. Notre modèle $BM25F\_S$ donne de meilleurs résultats en terme de MAP que le modèle simpliste $BM25\_S$.

Le tableau 4 confirme ce résultat par utilisateur, sur la collection $DelRIS$, pour les mesures MAP et $P[0.1]$.

	u_1		u_2	
	MAP	$P[0.1]$	MAP	$P[0.1]$
BM25	0.0683	0.0683	0.0339	0.0892
BM25_S	0.0651	0.0651	0.0406	**0.2176**
BM25F_S	**0.0706**	**0.0706**	**0.0451**	0.1610

TAB. 4 – *Détails utilisateur par utilisateur de l'évaluation avec la collection* $DelRIS$.

	MAP_S	$P_S[0.1]$
BM25	0,0461	0,0999
BM25_S	0,0382	0,1197
BM25F_S	**0,0627**	**0,1353**

TAB. 5 – *Evaluation avec la collection* $DelRIS2$.

Le tableau 5 présente les résultats des 3 modèles sur la collection $DelRIS2$. Cette fois-ci, avec cette collection que nous pensons être plus représentative du besoin de RIS sur le Web, on constate que le modèle $BM25\_S$ donne de moins bons résultats en terme de MAP_S que le modèle de référence $BM25$. Par contre, le modèle $BM25F\_S$ obtient les meilleurs résultats pour les deux mesures MAP_S et $P_S[0.1]$.

Le tableau 6 présente les résultats requête par requête, sur la collection $DelRIS2$, pour les mesures AP et $P[0.1]$.

u_x	$q_{l.u_x}$	***BM25***		***BM25_S***		***BM25F_S***	
		AP	$P[0.1]$	AP	$P[0.1]$	AP	$P[0.1]$
u_3	$q_{1.3}$	0.0475	0.0915	0.0390	0.0722	**0.0552**	**0.1111**
	$q_{2.3}$	0.0641	0.1061	0.0572	0.0968	**0.0821**	**0.1591**
	$q_{3.3}$	0.0410	0.0941	0.0239	0.0806	**0.1057**	**0.2105**
	$q_{4.3}$	0.0729	0.1111	0.0442	0.1026	**0.1003**	**0.1481**
	$q_{5.3}$	0.0559	0.1060	0.0402	0.0671	**0.0674**	**0.1184**
	$q_{6.3}$	0.0484	0.0921	0.0317	0.1053	**0.0778**	**0.1081**
	Moyenne	0.0550	0.1002	0.0394	0.0874	**0.0816**	**0.1426**
u_4	$q_{7.4}$	0.0036	0.0070	0.0010	0.0047	**0.0040**	**0.0098**
	$q_{8.4}$	0.0383	0.1304	0.0402	0.1250	**0.0514**	**0.1429**
	$q_{9.4}$	**0.0042**	0.0127	0.0035	**0.0215**	0.0025	0.0124
	$q_{10.4}$	0.0958	0.2000	0.0973	**0.4000**	**0.0999**	0.2857
	$q_{11.4}$	0.0358	0.1481	0.0416	**0.3333**	**0.0432**	0.1818
	Moyenne	0.0355	0.0997	0.0367	**0.1769**	**0.0402**	0.1265

TAB. 6 – *Détails requête par requête de l'évaluation avec la collection $DelRIS2$.*

On constate que le modèle $BM25F\_S$ obtient les meilleurs résultats pour les mesures AP et $P[0.1]$ pour toutes les requêtes, sauf pour la requête $q_{9.4}$.

L'intégration du contexte social de l'utilisateur *a priori* en utilisant le modèle d'indexation $BM25F\_S$ permet d'améliorer les résultats du système de RIS. L'amélioration est plus importante en terme de rappel (AP) que de précision ($P[0.1]$).

6 Conclusion

Cet article présente un modèle de Recherche d'Information Sociale (RIS) qui intègre le contexte informationnel social des utilisateurs, construit à partir de ses annotations sociales. Dans la première variante $BM25\_S$, le contexte social est intégré à la phase d'indexation sans équilibrer l'importance du contexte par rapport au contenu. Dans notre deuxième variante $BM25F\_S$, ce contexte est intégré par combinaison de termes à la phase d'indexation des ressources de manière *a priori*, en considérant un paramètre α permettant de régler son importance par rapport au contenu des ressources. L'indexation est donc personnalisée pour chaque utilisateur. Nous avons évalué nos modèles de RIS sur une collection de test de RIS construite à partir des annotations sociales du réseau *Delicious* et incluant des requêtes et jugements de per-

tinence centrée-utilisateur. Les résultats d'évaluation des modèles de RIS proposés montrent une amélioration par rapport au modèle de référence $BM25$.

Dans nos travaux futurs, nous envisageons d'exploiter les relations sociales au sein des RS pour enrichir le contexte informationnel social de l'utilisateur avec les annotations de son voisinage, en plus de ses annotations. Nous souhaitons aussi explorer les différentes méthodes de combinaison de score social pour pouvoir comparer et approfondir les résultats obtenus, et enfin expérimenter le modèle de RIS proposé sur d'autres collections de test, afin d'étudier sa robustesse et le passage à l'échelle.

Remerciements : Ce travail est partiellement soutenu par Saint-Étienne Métropole.

Références

Bao, S., G. Xue, X. Wu, Y. Yu, B. Fei, et Z. Su (2007). Optimizing web search using social annotations. In *Proceedings of the 16th conference on World Wide Web*, WWW'07, pp. 501–510. ACM.

Ben Jabeur, L., L. Tamine, et M. Boughanem (2010). A social model for literature access : towards a weighted social network of authors. In *Proceedings of the 9th conference Recherche d'Information Assistée par Ordinateur*, RIAO'10, pp. 32–39.

Ferrara, F. et C. Tasso (2011). Improving collaborative filtering in social tagging systems. In *Proceedings of the 14th Conference on Advances in Artificial Intelligence : spanish association for artificial intelligence*, CAEPIA'11, pp. 463–472.

Fischer, E. et A. R. Reuber (2010). Social interaction via new social media : (How) can interactions on twitter affect effectual thinking and behavior ? *Journal of Business Venturing 26*, 1–18.

Géry, M. et C. Largeron (2012). BM25t : a BM25 extension for focused information retrieval. *Knowledge and Information Systems 32*(1), 217–241.

Guy, I., N. Zwerdling, I. Ronen, D. Carmel, et E. Uziel (2010). Social media recommendation based on people and tags. In *Proceedings of the 33rd conference on Research and development in Information Retrieval*, SIGIR'10, pp. 194–201. ACM.

Jack, K., M. Hristakeva, et R. Garcia de Zuniga (2012). Mendeley's open data for science and learning : A reply to the DataTEL challenge | mendeley. *Special issue of Datasets and Data Supported Learning in Journal of Technology Enhanced Learning 4*, 31–46.

Kirsch, S. M. (2005). *Social Information Retrieval*. Ph. D. thesis, Rheinische Friedrich-Wilhelms-Universitat Bonn.

Konstas, I., V. Stathopoulos, et J. M. Jose (2009). On social networks and collaborative recommendation. In *Proceedings of the 32nd conference on Research and development in Information Retrieval*, SIGIR'09, pp. 195–202. ACM.

Li, B. et E. Gaussier (2012). Modèles d'information pour la recherche multilingue. In *COnference en Recherche d'Infomations et Applications*, CORIA'12, pp. 9–24.

Lin, Y., H. Lin, S. Jin, et Z. Ye (2011). Social annotation in query expansion. In *Proceedings of the 34th conference on Research and development in Information Retrieval*, SIGIR'11, pp. 405–414. ACM.

Manning, C. D., P. Raghavan, et H. Schutze (2008). *Introduction to Information Retrieval*. Cambridge University Press; 1st edition.

Robertson, S., H. Zaragoza, et M. Taylor (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th Conference on Information and Knowledge Management*, CIKM'04, pp. 42–49. ACM.

Robertson, S. E. et S. Walker (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th conference on Research and development in Information Retrieval*, SIGIR'94, pp. 232–241. ACM.

Schenkel, R., T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, et G. Weikum (2008). Efficient top-k querying over social-tagging networks. In *Proceedings of the 31st conference on Research and development in Information Retrieval*, SIGIR'08, pp. 523–530. ACM.

Theobald, M., R. Schenkel, et G. Weikum (2005). Efficient and self-tuning incremental query expansion for top-k query processing. In *Proceedings of the 28th conference on Research and development in Information Retrieval*, SIGIR'05, pp. 242–249. ACM.

Vallet, D., I. Cantador, et J. M. Jose (2010). Personalizing web search with folksonomy-based user and document profiles. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval*, ECIR'10, pp. 420–431. Springer Berlin Heidelberg.

Wen, K., R. Li, J. Xia, et X. Gu (2012). Optimizing ranking method using social annotations based on language model. *Artificial Intelligence Review 37*, 1–16.

Zaragoza, H., N. Craswell, M. Taylor, S. Saria, et S. Robertson (2004). Microsoft cambridge at TREC 13: Web and hard tracks. In *TExt Retrieval Conference*, TREC'04.

Summary

The Web has been heavily affected by recent advancements of Web 2.0 technologies, most remarkably by the emergence of social networks (SNs). In order to meet better Web users' needs, information retrieval (IR) must adapt its models and evaluation benchmarks in order to take into account the new patterns of information sharing enabled by SNs. The main challenges of social IR (SIR) are related to: (i) the definition of models able to describe the information needs of Web users in the context of SNs (SIR models); (ii) the representation of such models; and (iii) their evaluation. The last task is especially difficult because of the current lack of a comprehensive SIR data set and dedicated competitions. In this paper, we present a SIR model encompassing the social context of Web users within their SNs. We then evaluate this model over a SIR data set, built using the annotations of the collaborative bookmarking SN "del.icio.us".