Une Nouvelle Approche Sociale d'Expansion de Requêtes dans le Web 2.0

Mohamed Reda Bouadjenek[1,2], Hakim Hacid[1], Mokrane Bouzeghoub[2], Johann Daigremont[1]

[1]*Bell Labs, Alcatel-Lucent, France*
Centre de Villarceaux
91620, Nozay France
{Reda.Bouadjenek, Hakim.Hacid, Johann.Daigremont}@alcatel-lucent.com
[2]*Laboratoire PRiSM, Université de Versailles Saint Quentin-en-Yvelines*
45 Avenue des etats unis 78035, Versailles France
{Reda.Bouadjenek, Mokrane.Bouzeghoub}@prism.uvsq.fr

RÉSUMÉ. *Cet article aborde le problème d'expansion de requêtes qui consiste à enrichir les requêtes utilisateurs avec de l'information additionnelle pour maximiser son niveau de satisfaction en prenant en considération son écosystème. Tout en considérant les systèmes de bookmarking sociaux, l'approche proposée considère : (i) la similarité sémantique entre les termes qui composent les requêtes, (ii) la proximité sociale entre les termes qui composent les requêtes et les profils utilisateurs construits sur la base des intérêts de l'utilisateur et de ses voisins sociaux, et (iii) a la volé, une stratégie pour enrichir les requêtes utilisateurs. Cette approche a été évaluée sur un ensemble de données de del.icio.us et les résultats obtenus montrent son intérêt.*

ABSTRACT. *In this paper we address the problem of queries expansion and its personalization which consists of enriching user queries with additional information to maximize her satisfaction according to, e.g., her interests and her social ecosystem. While focusing on tagging systems, the proposed approach considers (i) the semantic similarity between tags composing a query, (ii) a social proximity between the query and the user profile for personalized expansion, and (iii) on the fly, a strategy for expanding user queries. The proposed approach has been evaluated using a large dataset crawled from del.icio.us and compared to the closest related work and the results show a clear superiority of our approach and promising results.*

MOTS-CLÉS : *Personalisation, Recherche d'Information Sociale, Réseaux Sociaux.*

KEYWORDS: *Personalization, Social Information Retrieval, Social networks.*

1. Introduction

Le Web 2.0, surtout avec sa dimension sociale, a introduit de nouvelles libertés pour l'utilisateur dans sa relation avec le Web. Ainsi, de l'information importante est généralement échangée sur ces plateformes dont on ne tire pas nécessairement profit en raison de la dynamique de cet écosystème, de l'énorme quantité volatile de l'information, etc. Dans ce contexte, il est naturel que l'un des besoins des utilisateurs les plus récurrents est de trouver de l'information pertinente dans cette énorme masse de données. Cependant, trouver des informations pertinentes devient encore plus difficile pour les utilisateurs finaux pour les raisons suivantes : (i) par définition, l'utilisateur ne sait pas ce qu'il recherche jusqu'à ce qu'il le trouve, et (ii) même si l'utilisateur sait ce qu'il cherche, il ne sait pas forcément comment formuler la bonne requête pour le système (e.g., du moment qu'il ignore la stratégie d'indexation).

L'expansion de requêtes est alors une solution pour réduire l'impact de tels problèmes. Elle consiste principalement à enrichir la requête initiale de l'utilisateur avec de l'information supplémentaire, de façon à ce que le système propose des résultats plus appropriés pour satisfaire les besoins de ce dernier. Bien que les plates-formes sociales augmentent la complexité de la recherche d'information (RI) comme le volume et l'hétérogénéité des données disponibles, et la grande variété des "profils" utilisateurs et de leurs besoins par exemple, elles pourraient fournir une opportunité intéressante pour l'amélioration du processus de RI en intégrant la dimension sociale. Le problème adressé peut être formulé comme suit : soit un ensemble de ressources annotées avec des tags libres par les utilisateurs dans une plateforme de bookmarking. Pour une requête $Q = \{t_1, t_2, \cdots, t_m\}$, comment transformer Q en $Q' = \{t'_1, t'_2, \cdots, t'_l\}$ avec une connaissance minimum de la stratégie d'indexation utilisée tel que : (i) Q est inclue dans Q', (ii) les résultats obtenus avec Q sont inclus dans ceux obtenus avec Q', et enfin, (iii) les résultats obtenus avec Q' augmentent la précision des résultats et ne diminuent pas la satisfaction des utilisateurs.

Cet article est organisé comme suit : la Section 2 présente notre proposition d'expansion de requêtes à l'aide de systèmes de bookmarking. La Section 3 aborde les expérimentations réalisées. La Section 4 présente quelques travaux connexes. Enfin, nous concluons et donnons quelques améliorations possibles dans la Section 5.

2. Expansion sociale de requêtes

Les systèmes de bookmarking sociaux fournissent aux utilisateurs un moyen pour annoter des ressources sur le Web, i.e., vidéo, pages web, images. Ces systèmes impliquent trois types d'entités : Utilisateurs (U), Tags (T), et Ressources (R). Soit Φ un ensemble de types défini comme suit : $\Phi = \{\Phi | 1 \leq i \leq s, \forall i, j : t_i \neq t_j\}$, qui représente les précédents types. Un système de bookmarking social peut être considéré comme un graphe tripartite non orienté : $\Omega(V, E, W)$, où V est un ensemble de n nœuds représentant des entités. Etant donné que les entités sont attachées à des types, nous définissons la fonction $\tau : V \rightarrow \Phi$ qui retourne le type de chaque nœud que

nous notons $\tau(v_i)$, et qui retourne dans ce cas soit U, T, ou R (i.e., $s = 3$ dans notre cas). Nous utilisons V_u, V_t, et V_r pour désigner respectivement les nœuds de type *Utilisateurs*, *Tags*, et *Ressources*.

Ainsi, la condition liée aux arêtes est que les entités d'une arête donnée doivent être de différents types. Les différentes entités du système de bookmarking social, qui participent dans ce que nous appelons "Action de Tagging", et qui donne un sens à ces plates-formes, sont dénotées sous forme d'un ensemble de triplés (U, T, R). Nous définissons trois paramètres n_t, n_r, et n_u, qui représentent respectivement le nombre de tags, de ressources et d'utilisateurs. Nous notons M_{ru} la matrice d'association $n_r \times n_u$ entre les ressources et les utilisateurs. De la même manière, nous définissons M_{tr} la matrice d'association $n_t \times n_r$ entre les tags et les ressources, et M_{ut} la matrice d'association $n_u \times n_t$ entre les utilisateurs et les tags. Enfin, W représente l'ensemble des poids associés aux arêtes du graphe.

Notre approche est basée essentiellement sur la création et la maintenance du graphe de tags G_t construit à partir du graphe social. Dans la section suivante, nous introduisons notre approche qui permet de construire ce graphe de tags ainsi que notre approche d'expansion de requêtes.

2.1. *Création du graphe de tags* G_t

Le graphe tripartie Ω peut être réduit en trois graphe bipartite, dont nous n'exploitons que les graphes *Utilisateurs-Tags* et *Ressources-Tags*. Dans ces graphes, chaque arête est pondérée selon son interaction avec les tags [MIK 07]. Ensuite, nous calculons les similarités entre tags dans les deux graphes bipartites précédents, afin d'obtenir deux graphes de tags ($G_{tag-users}$ et $G_{tag-ressources}$), l'un basé sur une agrégation par rapport aux utilisateurs, l'autre basé sur une agrégation par rapport aux ressources.

Afin de calculer cette similarité, nous proposons une mesure basée sur la popularité ainsi que l'importance des entités dans le graphe Ω. Au lieu de proposer une nouvelle mesure pour déterminer la popularité d'un nœud dans un contexte social, nous exploitons une technique déjà existante, *Social Page Rank (SPR)* [BAO 07]. *SPR* est initialement proposée pour calculer la popularité des ressources (i.e., pages Web) dans un système de bookmarking. L'idée principale derrière cet algorithme est la relation mutuelle existante entre les ressources populaires, les utilisateurs actifs et les tags fréquemment utilisés.

Etant donné que notre approche repose sur la popularité associée à chaque entité du graphe social, nous étendons *SPR* afin de calculer la popularité de tous les nœuds du graphe social (pas seulement ceux des documents). Dans le reste de cet article, et afin de distinguer les deux méthodes en fonction de leurs outputs, nous appelons cet algorithme *Social Entity Rank (SER)*. L'algorithme fournit trois vecteurs en sortie, à savoir : R, U et T représentant respectivement les popularités des ressources, des

utilisateurs et des tags. Enfin, l'equation 1 permet de calculer les similarités entres tags dans les deux graphes $G_{tag-users}$ et $G_{tag-ressources}$.

$$L_e(t_i, t_j) = \frac{\sum_{e \in N(t_i) \cap N(t_j)} Min(\omega(t_i, e), \omega(t_j, e)) \times Cred(e)}{\sum_{e \in N(t_i) \cup N(t_j)} Min(\omega(t_i, e), \omega(t_j, e)) \times Cred(e)} \quad (1)$$

La dernière étape consiste a fusionner (union) les deux graphes précédents, en calculant les similarités de la façon suivante :

$$Sim(t_i, t_j) = \alpha \times L_{documents} + (1 - \alpha) \times L_{users} \quad (2)$$

où α représente l'importance que l'on veut donner aux deux graphes de tags. Ce calcul de similarité fonctionne de la même manière que l'indice de *Jaccard* puisque c'est un rapport entre le taux des nœuds en commun de deux nœuds et l'union de leur nœuds voisins.

2.2. *Expansion effective de la requête*

Nous devons tout d'abord déterminer le profil de l'utilisateur avant d'effectuer l'expansion de la requête. En effet, un profil se compose principalement d'informations sur les centres d'intérêts de l'utilisateur construit en : (i) considérant les activités d'annotations de l'utilisateur, i.e., tags qu'il a utilisé, et (ii) les relations sociales (déduites) que l'utilisateur a avec d'autres utilisateurs, i.e., les tags que ces utilisateurs ont utilisé. L'ensemble de tags que l'utilisateur a utilisé est alors défini comme un vecteur $P^0_{ui} = (c_0, \cdots, c_{nt})$ pondéré de taille n, où $c_0, \cdots, c_{nt}$ correspondent au nombre de fois qu'un utilisateur u_i a utilisé le tag t_j.

Nous procédons ensuite à l'enrichissement de ce profil avec des informations provenant de son entourage social. Nous notons le nouveau profil avec P^*_{ui} calculé comme suit : $P^*_{u_i} = P^0_{u_i} + \sum_{u_j \in N(u_i)} Sim(u_i, u_j) \times P^0_{u_j}$ Par conséquent, nous donnons une plus grande importance et un plus grand poids aux tags qui sont plus utilisés par l'utilisateur u_i que les autres tags tout en augmentant l'importance des tags utilisés par des utilisateurs qui lui sont socialement similaires.

Enfin, le but de notre approche est de déterminer les voisins directes du terme initial de la requête $Q = t_0$, notés $N(t_0) = \{t_{01}, ..., t_{0n}\}$, et ensuite de calculer une distance sémantique entre chaque nœud de $N(t_0)$ et le profil de l'utilisateur. Notons que chaque tag de $N(t_0)$ est pénalisé en fonction de sa similarité avec le tag initial t_0. Cette distance est calculée en utilisant l'algorithme de *Diskjtra*, où au lieu de calculer le plus court chemin, nous calculons plutôt le chemin qui maximise la similarité. Les tags sont ensuite triés et sélectionnés en fonction de la plus grande similarité obtenue.

3. Evaluations préliminaires

Effectuer des évaluations pour la recherche personnalisée est un grand défi puisque le jugement de la pertinence des résultats ne peut être donné que par les utilisateurs eux-mêmes, ce qui est difficile à réaliser sur une grande échelle de données. Comme une évaluation primaire, nous avons opté pour une évaluation formelle suivant une approche similaire à celle proposée dans [CAR 09]. Ainsi, chaque bookmark (v_u, v_r, v_t) qui représente un utilisateur v_u qui a annoté un document v_d avec le tag v_t, peut être utilisé comme une requête de test pour l'évaluation. L'idée principale est basée sur l'hypothèse suivante : *pour une requête personnalisée $Q = (v_u, v_t)$ émise par l'utilisateur v_u avec le terme v_t, les documents pertinents sont ceux tagués par l'utilisateur v_u avec le tag v_t*. Les deux paramètres étudiés sont α (voir équation 2) qui permet de donner une plus grande importance soit aux documents ou aux utilisateurs, dans le cas du calcul de la similarité entre tags. Le deuxième paramètre est le nombre de tags avec lequel on réalise l'expansion.

Configuration des évaluations : Il existe de nombreux sites de bookmarking sociaux tel que *Del.icio.us* duquel nous avons extrait ses flux *RSS* durant Septembre 2010. L'ensemble de données obtenu se compose d'environ 300,000 posts, soit $15,473$ tags, $11,889$ utilisateurs, et $14,512$ pages Web (i.e., liens).

Dans cette étude hors ligne, pour chaque valeur de α, nous sélectionnons 100 couples (v_u, v_t) (v_t est le tag initial de la requête t_0), lesquels sont considérés comme un ensemble de requêtes personnalisées. Pour chaque couple, et selon l'hypothèse précédente, l'utilisateur v_u émet la requête $Q_0 = t$, où les documents pertinents sont ceux annotés par v_u à l'aide des tags de Q. Nous avons également supposé que les documents du graphe social sont indexés en fonction de leurs propres tags où leurs vecteurs sont pondérés en utilisant la mesure *TF-IDF*.

En utilisant le modèle vectoriel, i.e., pour représenter à la fois la requête et les documents dans le système, la similitude est calculée en utilisant la mesure du *cosinus*. Enfin, les documents sont classés en fonction de la valeur accordée par la mesure du *cosinus*. À partir de la liste de documents obtenue, l'*average-precision* sur les dix premiers résultats (AP@10) en considérant que tous les documents tagués par l'utilisateur v_u avec les tags inclus dans Q comme étant pertinents pour la requête. Comme dernière étape, le Mean-AP@10 est calculé sur les 100 requêtes. Enfin, la précision et le rappel sont calculés pour les TOP-5, TOP-10, TOP-15, TOP-20, TOP-25, TOP-30, TOP-35, et TOP-40 résultats, moyennés sur les 100 requêtes.

Impact des utilisateurs et des ressources sur l'expansion : L'importance des utilisateurs et des ressources sur la façon dont l'expansion est réalisée peut être réglée par le paramètre α de la Formule 1. En fixant $\alpha = 0$, on construit un graphe de tags (G_t) basé uniquement sur les usagers, alors qu'en fixant $\alpha = 1$ on construit un graphe de tags (G_t) basé uniquement sur les ressources communes (à savoir, les documents).

Les résultats obtenus montrent que le graphe de tags basés uniquement sur des liens utilisateurs donnent de meilleurs résultats que celui construits sur des liens de

ressources, ce qui confirme les résultats obtenus par Mika [MIK 07]. Ceci peut être expliqué par le fait que les utilisateurs qui ont les mêmes centres d'intérêts sont plus susceptibles d'utiliser les mêmes tags que les documents populaires auxquels les utilisateurs attribuent n'importe quels tags. Il convient de noter que notre approche améliore près de deux fois la qualité des résultats, en retrouvant plus de documents pertinents que l'utilisation de requêtes non-enrichies comme illustré dans la figure 1.

Impact du nombre de tags dans les requêtes : Nous évaluons un autre paramètre qui représente le nombre de tags qui peuvent enrichir la requête. L'objectif est de vérifier si la longueur d'une requête peut influer sur les résultats obtenus ou non. Le principe de cette évaluation est de faire varier à la fois α et le nombre de termes relatifs à $Q = t_0$. Notons que les requêtes non enrichies sont représentées par 1 sur l'axe des y dans la Figure 1.

La performance maximale est atteinte en ayant 3 à 6 termes liés au terme initial $Q = t_0$, et ce pour presque toutes les valeurs de α. Ajouter plus de sept termes n'a aucun effet, ou souvent a un impact négatif sur les performances pour les mêmes valeurs de α. Les résultats obtenus montrent également qu'il n'y a pas de corrélation avec la taille des requêtes puisque les meilleurs résultats sont obtenus avec $\alpha = 0$. Enfin, la meilleure taille pour les requêtes est de 4 selon les résultats.

Taux de requêtes sans documents pertinents : La Figure 2 illustre le taux de requêtes qui ne contiennent pas dans leurs **N** premiers résultats des documents pertinents (**N** est représenté par l'axe des X). Il est montré que notre approche (**dénoté SocExQ**) améliore de presque 10% les performances par rapport à l'utilisation de requêtes non enrichies (**Free**), ou des requêtes enrichies de façon brute en considérant uniquement les similarités entre termes (**ExpQ**). Ces résultats sont obtenus en fixant $\alpha = 0$ et une taille de requêtes de 4 termes.

Courbe precision-rappel : Enfin, les courbes de la Figure 3 montrent une comparaison entre notre approche et une approche par enrichissement brute. Il ressort que notre approche (**dénoté SocExQ**) améliore nettement la précision ainsi que le rappel par rapport à une approche brute centralisée autour du terme initial (**ExpQ**). Ces résultats sont aussi obtenus en fixant $\alpha = 0$ et une taille de requêtes de 4 termes.

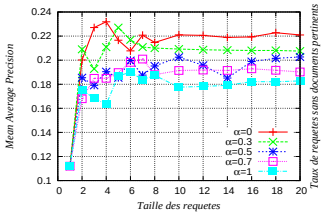

Figure 1. *MAP@10 pour les différentes valeurs de α, moyennées sur 100 requêtes pour différentes valeurs du nuage de tags. L'axe x représente le nombre de tags de la requête.*

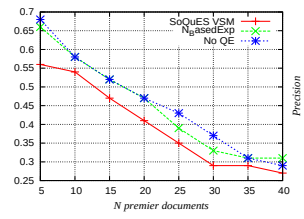

Figure 2. *Taux de requêtes sans documents pertinents dans leurs N premiers résultats.*

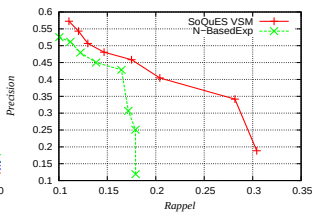

Figure 3. *Courbe precision vs rappel*

En résumé, l'évaluation hors ligne a montré que la meilleure performance est obtenue pour 3 à 6 termes ajoutés à la requête pour un graphe de tags basé uniquement sur des liens utilisateur. Les résultats obtenus sont deux fois mieux lorsqu'on utilise cette stratégie d'expansion de requêtes. Toutefois, il convient de noter que cette évaluation fournit des résultats qui ne donnent pas nécessairement une véritable vision des performances, en particulier à cause de la subjectivité élevée liée à ce type de systèmes. Ces résultats doivent être validés et renforcés avec l'aide d'utilisateurs finaux qui peuvent fournir une estimation plus précise du taux d'amélioration d'une telle approche. Nous avons commencé à mettre en place un tel protocole d'évaluation orienté utilisateurs finaux. Cette évaluation est basée sur un système de bookmarking interne appelé *People And Project (P&P)* au sein d'Alcatel-Lucent Bell Labs.

4. Travaux connexes

Plusieurs travaux ont été réalisés pour contribuer à la RI-S à partir de différentes perspectives. Bender et al. [BEN 08] ont considéré la RI-S, à la fois par l'expansion de requêtes et le classement des résultats, et ont proposé un modèle qui traite plus avec le classement des résultats qu'avec l'expansion de requêtes. En effet, les auteurs ont proposé un modèle de classement qui exploite directement les relations sociales en combinant des facteurs sémantiques et sociaux.

Dans le même esprit, Carmel et al. [CAR 09] proposent une approche basée sur la génération dynamique de profils. Cette approche intervient principalement sur la phase de classement des résultats avec l'introduction d'un paramètre social calculé en fonction des activités sociales d'un utilisateur. Pour y parvenir, plusieurs relations sociales sont prises en compte : (i) les relations de familiarité, (ii) les relations de similarité, (iii) les relations de centres d'intérêts, et (iv) toute autre relation. Sur la base de ces relations, un profil dynamique est construit et les résultats sont classés en fonction de ce profil. Notre travail est différent de cette proposition dans le sens où nous intervenons sur la réécriture des requêtes.

Une autre vision de l'introduction de la dimension sociale dans le processus de RI est la combinaison des résultats de recherche Web avec des informations sociales afin de re-classer les résultats selon les affinités sociales, à savoir l'amitié sociale. Le travail de Noll et Meinel [NOL 07] est un exemple de telles initiatives qui repose sur deux types de profils : (i) le profil de l'utilisateur et (ii) le profil du document. L'approche proposée combine deux principaux services pour obtenir des résultats : (i) Un service de recherche qui permet d'obtenir une liste de documents qui correspondent à une requête, (ii) un service de bookmarking qui permet d'obtenir les profils des documents récupérés. Un profil de l'utilisateur est également construit et mixé avec les deux structures déjà construites de manière à réaliser une stratégie de reclassement des documents afin de matcher avec les profils des utilisateurs.

5. Conclusion et perspectives

Cet article propose une contribution à la partie expansion de requêtes dans la recherche d'information. Nous avons proposé une nouvelle approche basée sur la dimension sociale afin de transformer une requête initiale Q en une autre requête Q' enrichie avec des termes fortement liés aux termes initiaux de la requête. Cette approche a été intégrée dans un système appelé **SoQuES** qui peut être facilement connecté sur des plates-formes de bookmarking sociales existantes. Enfin, une évaluation formelle de la qualité des résultats et du comportement des différents paramètres, en utilisant un ensemble de données extrait à partir de *del.icio.us*, a montré les avantages de cette approche.

Comme objectif à court terme, nous avons l'intention de valider cette approche en utilisant un sondage à grande échelle auprès des utilisateurs en utilisant notre système de bookmarking social interne *P&P*. À long terme, nous espérons améliorer le temps d'exécution de la génération des requêtes en réduisant la complexité algorithmique. Nous sommes aussi convaincus que la combinaison de notre approche avec des fonctions sociales de classement de documents telles que celles proposées dans [CAR 09, XU 08, NOL 07, HOT 06, BEN 08] peuvent être d'un grand intérêt. Enfin, nous développons actuellement un prototype stable qui peut être embarqué sur le moteur de recherche du système interne de bookmarking (P&P) pour tester les performances dans un environnement réel.

6. Bibliographie

[BAO 07] BAO S., XUE G.-R., WU X., YU Y., FEI B., SU Z., « Optimizing web search using social annotations », *WWW*, 2007, p. 501-510.

[BEN 08] BENDER M., CRECELIUS T., KACIMI M., MICHEL S., NEUMANN T., PARREIRA J. X., SCHENKEL R., WEIKUM G., « Exploiting social relations for query expansion and result ranking. », *ICDE Workshops*, IEEE Computer Society, 2008, p. 501-506.

[CAR 09] CARMEL D., ZWERDLING N., GUY I., OFEK-KOIFMAN S., HAR'EL N., RONEN I., UZIEL E., YOGEV S., CHERNOV S., « Personalized social search based on the user's social network », *CIKM*, New York, USA, 2009, ACM, p. 1227–1236.

[HOT 06] HOTHO A., JÄSCHKE R., SCHMITZ C., STUMME G., « Information Retrieval in Folksonomies : Search and Ranking », *The Semantic Web : Research and Applications*, , 2006, p. 411–426.

[MIK 07] MIKA P., « Ontologies are us : A unified model of social networks and semantics », *Web Semant.*, vol. 5, n° 1, 2007, p. 5–15, Elsevier Science Publishers B. V.

[NOL 07] NOLL M. G., MEINEL C., « Web Search Personalization Via Social Bookmarking and Tagging », vol. 4825, 2007, p. 367–380, Springer.

[XU 08] XU S., BAO S., FEI B., SU Z., YU Y., « Exploring folksonomy for personalized search », *SIGIR*, New York, USA, 2008, ACM, p. 155–162.