Exploitation des contributions des usagers liées au *social bookmarking* pour améliorer la Recherche d'Information

Mădălina Mitran

Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9

mitran@irit.fr

RÉSUMÉ. *Les moteurs de recherche communs exploitent le contenu des documents qu'ils indexent. Or, les internautes créent également des données explicites (*tags*, annotations, commentaires, notes, données de géoréférencement, etc.) et implicites (clics, logs, etc.) qu'il semble utile de prendre en compte pour améliorer l'indexation. Nos travaux concernent actuellement deux problématiques. Premièrement, comment analyser les bookmarks sociaux pour en extraire les centres d'intérêts des individus et leurs tendances tout en prenant en compte la dimension temporelle? Cette analyse permettrait de mieux les modéliser pour personnaliser leur recherche d'information ou leur recommander de l'information, par exemple. Deuxièmement, comment indexer des photos géoréférencées présentes sur des plateformes de partage afin de mieux les valoriser?*

ABSTRACT. *Common search engines index documents according to their contents. Despite this, users create explicit (e.g., tags, annotations, comments, ratings, location data) and implicit data (e.g., clicks, logs) that seem useful to take into account to improve the indexing process. Our current works addresses two issues. First, how to analyze social bookmarks to extract people's interests and their tendencies while taking into account the temporal dimension? This analysis allows for example to better model the users' interests in order to personalize search, as well as to recommend information to them. Second, how to index georeferenced pictures present on collaborative platforms to better value them?*

MOTS-CLÉS : *web participatif,* social bookmarking*, tags, indexation, photos, localisation*

KEYWORDS: *web 2.0, social bookmarking, tags, indexing, photos, localization*

Mădălina Mitran

1. Introduction

En Recherche d'Information (RI), les documents sont classiquement indexés en fonction de leur contenu, qu'il soit textuel ou multimédia. Les moteurs de recherche, qui s'appuient sur ces index, sont aujourd'hui des outils performants, répandus et indispensables. Ils visent à fournir des réponses pertinentes selon le besoin de l'utilisateur. Dans le web initial des années 1990, l'individu était un simple consommateur d'information. L'évolution du web vers le web 2.0, également appelé web participatif, permet désormais aux internautes d'être producteurs de ressources, en plus d'être consommateurs. En effet, ils peuvent publier des textes (un billet dans un blog, par exemple), des photos, des vidéos, etc. qui sont diffusés au travers de diverses plateformes. Parmi ces plateformes, on trouve notamment : ① les encyclopédies participatives, telles que `wikipedia.org` ; ② les plateformes de blog, telles que `blogger.com` ; ③ les plateformes de *social bookmarking*, telles que `delicious.com`, `connotea.org` soutenu par la maison d'édition *Nature* ou IBM DogEar (Millen *et al.*, 2006). Elles permettent à un utilisateur de conserver des ressources identifiées par une URL et annotées par des « tags » (expressions librement choisies par les individus) ; ④ les plateformes de partage de photos et vidéos, telles que `flickr.com`, `panoramio.com` ou `youtube.com` et ⑤ les plateformes de réseaux sociaux, telles que `facebook.com`.

Dans ce contexte du web 2.0, une multitude de données additionnelles sont créées par les internautes. Ce sont notamment des « traces » qu'ils associent aux documents : des jugements de pertinence (votes), des *bookmarks*, des *tags*, des commentaires, des annotations, des débats argumentatifs, des descriptions, etc. De toute évidence, ces données contribuées spontanément par de multiples individus indépendants reflètent des aspects complémentaires au seul contenu des documents. Pourtant, ces aspects ne sont pas ou peu intégrés au processus de RI. Il s'agirait alors de trouver des moyens pour modéliser puis exploiter toutes ces données accessibles dans le but d'améliorer la RI. Nos travaux de thèse abordent cette problématique au travers de deux contextes. Premièrement, en analysant l'activité d'un usager sur une plateforme de *social bookmarking* telle que `connotea.org`. Deuxièmement, en exploitant les *tags* associés aux photos géoréférencées publiées sur une plateforme de partage telle que `flickr.com`.

Cet article est organisé comme suit : dans la section 2, nous présentons les approches d'état de l'art liées à la valorisation des *tags*, afin d'améliorer la RI. La section 3 détaille nos contributions pour les deux contextes mentionnés précédemment. Enfin, nous concluons en section 4 où nous détaillons également les perspectives à ces travaux.

2. Exploitation du web participatif : approches de la littérature

Un système de RI peut relever de deux paradigmes : le *pull* et le *push* (Belkin *et al.*, 1992). Le *pull* représente le fait que l'utilisateur soumet une requête à un moteur de recherche, pour retrouver des documents pertinents par rapport à son besoin. Le *push* représente l'automatisation de la recherche : l'utilisateur reçoit automatiquement des recommandations que le système juge pertinentes pour lui. Ces deux paradigmes peuvent être améliorés en prenant en compte le profil de l'utilisateur ou des données issues du web participatif. Nous considérons en particulier les données de type *tags*

issues des plateformes de *social bookmaring* qui ont connu une évolution rapide ces dernières années. Des individus utilisent de tels outils pour sauvegarder et (re)trouver des ressources. Golder *et al.* (2006) ont analysé l'utilisation des *tags*, les activités des individus et le type de ressources annotées. Par ailleurs, Bischoff *et al.* (2008) ont effectué une analyse des systèmes comme `delicious.com` et `last.fm`, pour identifier les *tags* utiles pour accéder à l'information. En complément à ces deux études, l'exploitation des *tags* s'est révélée fructueuse dans le cadre de la recommandation (Klasnja Milicevic *et al.*, 2010), pour la visualisation des *tags* en fonction du temps (Dubinko *et al.*, 2007) ou pour la recherche personnalisée qui est basée sur les profils des individus (Cai *et al.*, 2010). Une des limites de l'état de l'art concerne l'absence de modélisation des centres d'intérêts des utilisateurs en prenant en compte la dimension temporelle et les *tags*.

En complément aux ressources de type texte, nous nous sommes intéressés aux plateformes de *social bookmarking* dédiées à un autre média : les photos. La problématique de leur indexation a été traitée en fonction de leur contenu (Lai *et al.*, 1998) ou de leur contexte d'apparition dans un document (Rao *et al.*, 2007). De plus, le développement du web et la démocratisation des appareils photo numériques et mobiles (téléphone portable, pas exemple) produisent des metadonnées (localisation GPS, date et heure, auteur, etc.) qui pourraient être intégrés dans le processus d'indexation. Lee *et al.* (2010) proposent une méthode d'indexation automatique des images en utilisant les informations de géoréférencement et de date/heure au moment de la capture pour faciliter la recherche ultérieure. Par contre, à notre connaissance l'exploitation conjointe des *tags* et des metadonnées n'a pas été réalisée.

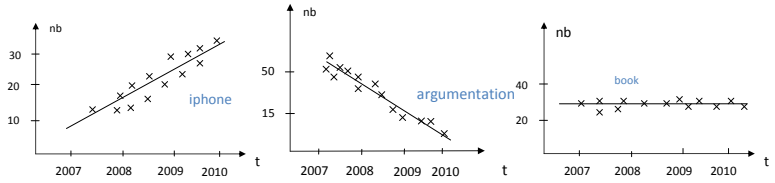
Les approches de la littérature présentées sont basées sur des données additionnelles que les usagers créent en utilisant des services sur le web. Certains aspects n'ont pas été abordés à notre connaissance, nous les détaillons dans la section suivante.

3. Propositions : améliorer *pull* et *push* dans le contexte du web participatif

3.1. *Représentations des activités des usagers du* social bookmarking

Nous avons proposé plusieurs approches qui visent à analyser les bookmarks sociaux pour en extraire les centres d'intérêts des individus et leurs tendances, en considérant la dimension temporelle (Mitran, 2010; Mitran *et al.*, 2010). Afin de répondre au problème de représentation des activités des individus, nous avons modélisé leur activité d'étiquetage : l'emploi des *tags* (nb) en fonction du temps (t). Pour chaque *tag*, nous calculons sa fréquence d'utilisation, puis sa tendance (coefficient directeur a) en calculant la régression linéaire $nb = a \cdot t + b$. Par exemple, la Figure 1 montre l'utilisation des *tags* « book », « argumentation », « iphone » d'un individu dans la période t. Cette représentation permet de voir l'évolution de ses centres d'intérêt.

On peut déduire plusieurs caractéristiques sur un usager à partir de la représentation de son activité en fonction du temps : ① ses *tags* les plus représentatifs ; ② ses intérêts actuels ; ③ ses intérêts sur le court et long terme, etc. De plus, on peut calculer les



(a) Intérêt croissant (b) Intérêt en déclin (c) Intérêt récurrent

Figure 1 – Exemple de trois tendances pour les intérêts d'un usager

similarités entre usagers, puis former des communautés qui s'intéressent aux mêmes intérêts sur la même période de temps. Cela permettra au système de mieux représenter ses usagers en fonction de leurs activités. Pour les individus, ils bénéficieront d'une meilleure vision des activités des autres usagers. Actuellement, nous concevons une évaluation pour confronter la perception des usagers sur leurs activités par rapport aux résultats de notre algorithme. Nous envisageons d'effectuer des calculs sur les données de `connotea.org` et de solliciter les usagers correspondants pour évaluer les résultats.

3.2. *Indexation des photos étiquetées et géoréférencées*

Dans cette section nous considérons le problème de l'indexation de photos en ignorant leur contenu et leur contexte d'utilisation (documents qui les contient). Ce cas correspond à l'indexation de photos collectives publiées sur les plateformes de partage, où l'on ne dispose que des photos avec leurs metadonnées de géoréférencement et des *tags* fournis par les utilisateurs. Pour une photo p, nous proposons donc une approche d'indexation représentée en Figure 2, constituée de trois étapes :

1) les pages de `wikipedia.org` contiennent des coordonnées de géoréférencement (latitude, longitude). Or, nous faisons l'hypothèse que la page `wikipedia.org` qui correspond à la localisation où p a été faite contient des termes descriptifs (les pages Wikipédia des lieux contiennent les coordonnées GPS associées). Nous utilisons des méthodes classiques de RI pour extraire ces termes (modèle vectoriel, pondération tf·idf, par exemple). La liste L_g représentée dans la Figure 2 est produite ;

2) nous considérons les *tags* que les individus ont mis pour décrire les photos $\{p_1, \ldots, p_n\}$ prises dans l'environnement immédiat de p (près de p en s'appuyant sur `panoramio.com`). L'environnement immédiat est déterminé par le calcul de distance entre p et $\{p_1, \ldots, p_n\}$ (indépendamment de l'échelle de la carte). La liste L_p représentée dans la Figure 2 est produite ;

3) nous combinons le résultat des deux premières approches. Ici nous considérons à la fois les termes descriptifs issus de `wikipedia.org` et les *tags* extraits à partir de la géoréférence. Pour ce faire, nous utilisons le combinateur CombMNZ (Fox *et al.*, 1994) normalisé, issu de travaux de RI. Il combine plusieurs listes de résultats (L_g et L_p ici) issues de différents approches en une seule liste (L_f).

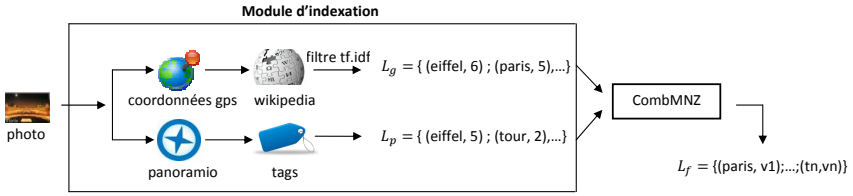


Figure 2 – Indexation d'une photo étiquetée et géoréférencée

Concernant l'évaluation de cette proposition, nous proposons de confronter les listes de termes (L_g, L_p, L_f) avec une indexation manuelle. Pour ce faire, nous établissons une liaison avec l'évaluation en RI où les résultats d'un système pour une requête sont comparés aux jugements de pertinence effectués par des humains. Cette comparaison est réalisée à l'aide des mesures de pertinence. Dans notre cas nous pouvons utiliser la mesure NDCG, *Normalized Discounted Cumulative Gain* (Järvelin *et al.*, 2002) qui prend en compte des jugements graduels (les termes d'indexation produits sont plus ou moins pertinents). De plus, à notre connaissance, il n'existe pas de collection de données (*CLEF*, *TREC*, *NTCIR* ou autre) adaptée pour l'évaluation de notre proposition. C'est pourquoi nous envisageons de construire une collection adaptée.

4. Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à la recherche d'information basée sur la participation des internautes liée à l'activité de *social bookmarking* et de partage de photos, notamment. Notre contribution est double. Premièrement, nous déterminons les centres d'intérêts des individus à partir de leurs *tags* en prenant en compte la dimension temporelle. Deuxièmement, nous proposons d'indexer les photos à partir des *tags* publiés et de leurs métadonnées de géoréférencement. Des limites peuvent être identifiées dans nos travaux. Par exemple, la régression linaire ne représente pas toujours de façon optimale les intérêts des utilisateurs. Par ailleurs, il existe des photos qui ne sont pas géoréférencées et des photos pour lesquelles il n'y a pas de page Wikipédia les décrivant. C'est le cas des photos d'objets ou de personnes, par exemple.

Une perspective à long terme pour notre travail est l'application de la théorie de Gladwell (2002) sur les plateformes de *social bookmarking*. Elle aborde la transmission de l'information au regard des théories du domaine de l'épidémiologie. L'information est assimilée à un virus qui est propagé différemment selon les individus. Selon Gladwell, trois catégories d'individus sont à considérer : les *connectors* sont caractérisés par leur grand nombre d'accointances, ils arrivent à établir des liens entre différentes communautés ce qui leur permet d'y disséminer l'information, les *mavens* accumulent les savoirs, disposent et sont à l'origine de nombreuses informations qu'ils partagent volontiers autour d'eux, dans un cercle réduit d'accointances et les *salesmen* promeuvent les nouvelles idées qu'ils glanent, savent les valoriser et les diffuser autour d'eux. Cette théorie pourrait être transposée dans le cadre expérimental du web participatif. En effet, l'observation des interactions entre individus et de leurs productions pourrait

alimenter un modèle d'interaction sociale. Puis, l'analyse de ces données permettrait d'identifier les caractéristiques des usagers (*connectors*, *mavens*, *salesmen*). Cette connaissance permettrait alors d'adapter le processus de RI en détectant les nouvelles informations et **tendances** (*mavens*), en les recommandant aux personnes les plus à même de les **valoriser** (*salesmen*) pour accroître leur **visibilité** et leur **dissémination** dans les différents cercles d'accointances (*connectors*).

5. Bibliographie

Belkin N. J., Croft W. B., « Information filtering and information retrieval : two sides of the same coin ? », *ACM*, vol. 35, n° 12, p. 29–38, 1992.

Bischoff K., Firan C. S., Nejdl W., Paiu R., « Can all tags be used for search ? », *CIKM*, ACM, p. 193-202, 2008.

Cai Y., Li Q., « Personalized search by tag-based user profile and resource profile in collaborative tagging systems », *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, New York, NY, USA, p. 969–978, 2010.

Dubinko M., Kumar R., Magnani J., Novak J., Raghavan P., Tomkins A., « Visualizing tags over time », *ACM Trans. Web*, vol. 1, n° 7, p. 1559-1131, August, 2007.

Fox E. A., Shaw J. A., « Combination of multiple searches », *The Second Text REtrieval Conference (TREC-2)*, NIST, p. 243-252, 1994.

Gladwell M., *The Tipping Point : How Little Things Can Make a Big Difference*, Back Bay Books, 2002.

Golder S. A., Huberman B. A., « Usage patterns of collaborative tagging systems », *Journal of Information Science*, vol. 32, n° 2, p. 198–208, 2006.

Järvelin K., Kekäläinen J., « Cumulated gain-based evaluation of IR techniques », *ACM Trans. Inf. Syst.*, vol. 20, p. 422–446, October, 2002.

Klasnja Milicevic A., Nanopoulos A., Ivanovic M., « Social Tagging in Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions », *Artificial Intelligence Review*, vol. 33, n° 3, p. 187-209, 2010.

Lai T.-S., Tait J., « Using Global Colour Features for General Photographic Image Indexing and Retrieval. », *SIGIR*, ACM, p. 349-350, June, 1998.

Lee Y.-H., Kim B., Kim H.-J., « Photograph Indexing and Retrieval using Combined Geo-information and Visual Features. », *CISIS*, IEEE Computer Society, p. 790-793, April, 2010.

Millen D. R., Feinberg J., Kerr B., « Dogear : Social bookmarking in the enterprise », *CHI '06 : Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, New York, p. 111–120, 2006.

Mitran M., Recherche d'information sociale : exploitation du social bookmarking pour enrichir l'accès à l'information, Rapport de master, Université Paul Sabatier, Toulouse, juin, 2010.

Mitran M., Cabanac G., Boughanem M., « Détection des intérêts et de leurs tendances pour des usagers sur des plateformes de social bookmarking (poster) », *VSST'10 : Colloque Veille Stratégique Scientifique et Technologique*, IRIT, octobre, 2010.

Rao N. G., Kumar V. V., « Texture based image indexing and retrieval. », *VISAPP (Special Sessions)*, INSTICC - Institute for Systems and Technologies of Information, Control and Communication, p. 177-181, Avril, 2007.