RI sociale: intégration de propriétés sociales dans un modèle de recherche

Ismail Badache

RI sociale : intégration de propriétés sociales dans un modèle de recherche

Ismail Badache[1]

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS, SIG
118 Route de Narbonne
F-31062 Toulouse cedex 9
France
ismail.badache@irit.fr

RÉSUMÉ. *Cet article propose une approche de recherche d'information, basée sur le contenu généré par l'utilisateur (CGU). Nos travaux se focalisent sur l'exploitation des CGUs dans la recherche des ressources web (pages, vidéos, etc). En particulier, nous nous intéressons à identifier, extraire et quantifier, à partir de plusieurs réseaux sociaux, certaines propriétés de ces CGUs, telles que la popularité et la confiance. Ces propriétés vont être intégrées dans un modèle de ranking. Plus précisément, nous proposons un modèle qui prend en considération ces propriétés sociales, en les combinant avec la pertinence thématique afin d'améliorer le tri des résultats renvoyés par un moteur de recherche. Nous avons évalué notre modèle sur une collection de test extraite du site Web "imdb.com". Les résultats obtenus montrent l'efficacité de notre modèle par rapport à la recherche d'information classique.*

ABSTRACT. *This paper proposes an information retrieval approach, based on user generated content (UGC). Our work focuses on the use of UGCs to seek Web resources (pages, videos, etc.). In particular, we are interested to identify, extract and quantify, from several social networks, some properties of these UGCs, such as the popularity and trust. These properties will be included in a ranking model. More precisely, we propose a model that takes into account these social properties, in combination with topical relevance to improve search engine returned results. We evaluated our model on a test collection extracted from "imdb.com" website. The obtained results show the effectiveness of our model compared to classical information retrieval.*

MOTS-CLÉS : *Contenu généré par l'utilisateur, recherche d'information sociale, réseaux sociaux, propriétés sociales.*

KEYWORDS: *User generated content, social information retrieval, social networks, social properties.*

1. Directeur de thèse : Mohand BOUGHANEM

1. Introduction

Le Web 2.0 a complètement changé la façon dont les personnes partagent des informations. Il permet aux utilisateurs d'interagir, produire et partager des masses importantes de contenus sociaux. Ces derniers, connus sous l'abréviation anglaise UGC (*User Generated Content*), prennent de plus en plus d'intérêt dans le Web.

Selon (Baeza-Yates, 2009): « *User Generated Content is one of the main current trends in the Web. This trend has allowed all people that can access the Internet to publish content in different media, such as text (e.g. blogs), photos or video* ».

Selon (Volkovich *et al.*, 2011): « *Social news websites have gained significant popularity over the last few years. The participants of such websites are not only allowed to share news links but also to annotate, to evaluate and to comment them* »

Ces contenus sociaux (tag, commentaire, mention, etc) ont conduit à l'émergence d'un nouveau type de recherche d'information (RI), appelé recherche d'information sociale (RIS). Selon (Karweg *et al.*, 2011): « *Social search is a variant of information retrieval where a document or website is considered relevant if individuals from the searcher's social network have interacted with it* ».

La problématique de la RIS est de répondre à plusieurs questions : (a) la première concerne le développement d'outils pour répondre aux besoins des utilisateurs : quels sont les besoins d'information des utilisateurs de médias sociaux ? Quels modèles de RIS ? (b) la seconde concerne l'exploitation des UGCs dans des tâches de RI : quelles propriétés sociales utiles peuvent être exploitées pour la recherche des ressources (pages web, vidéos, etc) ? Comment les extraire et les quantifier ?

Cet article présente dans la section 2 les travaux connexes. Notre approche est décrite en section 3. La section 4 est consacrée à l'évaluation de l'approche sur une collection issue du site 'imdb.com'. Enfin, la section 5 conclut l'article.

2. Travaux relatifs

Les travaux exploitant les UGCs peuvent être classés en deux catégories. La première tente d'intégrer l'UGC pour l'expansion de la requête. Tandis que la seconde catégorie de travaux tente d'intégrer ces UGCs dans la phase de tri des résultats. Nous nous intéressons dans cet article aux travaux appartenant à la deuxième catégorie. Ces travaux se focalisent sur deux axes : (a) le premier concerne l'identification de l'importance d'une ressource à travers ces UGCs; (b) le second prend, en plus, en considération l'importance sociale de l'auteur de la ressource.

Dans le premier axe, (Bao *et al.*, 2007) proposent deux algorithmes : *Social PageRank* et *Social SimRank*, pour calculer les scores, respectivement, de popularité des pages Web et la similarité requête-annotations. Ils montrent par la suite que la combinaison des algorithmes améliore de manière significative la précision moyenne.

(Hong *et al.*, 2011) exploitent le nombre de retweets comme mesure de popularité, au sein d'un classifieur pour prédire si de nouveaux messages seront retweetés et à

quelle fréquence. Cependant, des tweets banals (rumeurs, sans intérêt) peuvent être très populaires tels que ceux concernant des célébrités, qui possèdent généralement un grand nombre de followers. En revanche, notre approche exploite d'autres UGCs pour mieux mesurer la popularité ainsi que l'importance de l'information publiée.

Dans le second axe, (Karweg *et al.*, 2011) proposent une approche combinant un score thématique et un score social basé sur deux facteurs sociaux : (a) l'intensité d'engagement d'un utilisateur pendant une interaction avec un document, mesurée à partir du nombre de clics, votes, enregistrement et recommandation; (b) le degré de confiance pour chaque utilisateur mesuré à partir du graphe social, en utilisant l'algorithme de *PageRank*. (Pal *et al.*, 2011) proposent un modèle d'identification des auteurs les plus influents dans Twitter. Cette solution, utilisée en RI, est basée sur le modèle de mélange gaussien en exploitant des données sociales telles que : le nombre de retweets, tweet conversationnel et followers actifs par rapport au sujet d'intérêt. Par rapport à ces deux propositions, notre approche ne se limite pas qu'à la recherche dans Twitter et exploite des UGCs à travers plusieurs réseaux sociaux.

(Ben Jabeur *et al.*, 2011) proposent une approche de RIS dans les microblogs en exploitant les données sociales de Twitter. Ils présentent l'influence d'un bloggeur à travers les relations de rediffusion des tweets et exploitent cette influence ainsi que l'expertise du bloggeur pour la recherche des tweets. Plus précisément, leur modèle combine la pertinence thématique des publications avec un score d'influence sociale, obtenu par l'application de l'algorithme *PageRank* sur le réseau social de rediffusion, et un score d'expertise obtenu par un modèle de langue. Par rapport à notre approche, nous calculons le score social en fonction d'autres facteurs sociaux, comme l'importance et la popularité de l'information ainsi que sa fraicheur sociale.

3. Approche proposée

Notre approche consiste à exploiter ces UGCs afin d'améliorer le processus de tri des résultats des moteurs de recherche. Nous tentons d'identifier, extraire et quantifier certaines propriétés à partir de ces UGCs, puis de les intégrer dans un modèle de recherche combinant un score social et un score thématique.

3.1. *Propriétés sociales étudiées et leur quantification*

Nous avons identifié ces propriétés sociales en analysant différents UGCs à travers plusieurs réseaux sociaux mais plusieurs questions se posent : Comment les quantifier ? Quels UGCs doit-on exploiter ? Et à partir de quels réseaux sociaux ?

Les UGCs exploités, récupérés à travers les différentes APIs, de chaque réseau social et pour chaque propriété sont présentés dans le tableau suivant :

	Propriétés sociales	UGCs de quantification	Réseaux sociaux
Auteur	Importance et influence	Nombre de « *followers* »	Twitter
		Nombre de « *abonnés* » Nombre de « *amis* »	Facebook

Ressource	Popularité	Nombre de « *J'aime* » Nombre de « *Commentaire* »	Facebook
		Nombre de « *Tweet* »	Twitter
		Nombre de « *Partage* »	LinkedIn, Pinterest, Facebook
		Nombre de « +1 »	Google+
		Nombre de « *Marque* »	Delicious, StumbleUpon, Digg
		Nombre de « *Vote* »	Imdb
	Importance	Nombre de « +1 »	Google+
		Nombre de « *J'aime* »	Facebook
		Nombre de « *Vote* »	Imdb
	Fraicheur	Date de la dernière mention	Facebook

Tableau 1. *Les UGCs exploités dans la quantification des propriétés sociales*

3.2. *Formulation du modèle de recherche*

Les deux scores de pertinence thématique et sociale sont combinés linéairement selon la formule normalisée suivante :

$$SCORE\ (Q, r, G) = \alpha * SCORE_{Thématique}\ (Q, r) + (1- \alpha) * SCORE_{Social}\ (Q, l_r, G) \quad [1]$$

Avec Q, r, G représentent respectivement la requête, la ressource et le réseau social. $\alpha \in [0,1]$ un paramètre de pondération. $SCORE_{Thématique}\ (Q, r)$ est le score normalisé de la pertinence thématique. $SCORE_{Social}\ (Q, l_r, G)$ est le score normalisé de la pertinence sociales, avec l_r correspondant au lien URL de la ressource.

3.3.1. *Score thématique*

La pertinence thématique $SCORE_{Thématique}(Q, r)$ dépend des paramètres par défaut du système de recherche d'information *Lucene Solr* basé sur le modèle vectoriel.

3.3.2. *Score social*

Nous précisons que le score social $SCORE_{Social}\ (Q, l_r, G)$ est calculé à partir de ces propriétés sociales issues des réseaux sociaux G. Elles sont sous forme de deux facteurs combinés linéairement selon la formule normalisée [2] suivante :

$$SCORE_{Social}(Q, l_r, G) = \frac{\beta\ Fraicheur_{Sociale}(l_r) + (1-\beta)\ Pop\_Imp_{Sociale}(l_r)}{MAX_{l_r}(SCORE_{Social}(Q, l_r, G))} \quad [2]$$

Avec $\beta \in [0,1]$ un paramètre de pondération.

– **$Fraicheur_{Sociale}$:** la fraicheur est mesurée par rapport à la date de la dernière mention envers la ressource sur le réseau social. La formule normalisée [3] de ce facteur est donnée comme suit :

$$Fraicheur_{Sociale}(l_r) = \frac{Temps_{Dernière\_mention}(l_r)}{MAX_{l_r}\left(Temps_{Dernière\_mention}(l_r)\right)} \quad [3]$$

– **$Pop\_Imp_{Sociale}$:** est l'indice de la popularité et l'importance d'une ressource. La formule qui combine les différents critères issus des différents réseaux sociaux est comme suit :

$$Pop\_Imp_{Social}(l_r) = \sum_{i=1}^{9} f_{Critère\ i}(l_r) \quad [4]$$

Avec :

$f_{Critère\ 1}(l_r) = f_{Facebook}(l_r) = Nbr_{J'aime}(l_r) + Nbr_{Partage}(l_r) + Nbr_{Commentaire}(l_r)$

$f_{Critère\ 2}(l_r) = Nbr_{Twitter\_Tweet}(l_r)$; $f_{Critère\ 3}(l_r) = Nbr_{LinkedIn\_Partage}(l_r)$

$f_{Critère\ 4}(l_r) = Nbr_{Pinterest\_Partage}(l_r)$; $f_{Critère\ 5}(l_r) = Nbr_{Google\_+1}(l_r)$

$f_{Critère\ 6}(l_r) = Nbr_{Delicious\_Marque}(l_r)$; $f_{Critère\ 7}(l_r) = Nbr_{StumbleUpon\_Marque}(l_r)$

$f_{Critère\ 8}(l_r) = Nbr_{Digg\_Marque}(l_r)$; $f_{Critère\ 9}(l_r) = Nbr_{IMdb\_Vote}(l_r)$

La normalisation de la formule [4] est comme suit :

$$Pop\_Imp_{Sociales}(l_r) = \frac{Pop\_Imp_{Sociale}(l_r) - MIN_{l_r}(Pop\_Imp_{Sociale}(l_r))}{MAX_{l_r}(Pop\_Imp_{Sociale}(l_r)) - MIN_{l_r}(Pop\_Imp_{Sociale}(l_r))} \quad [5]$$

4. Evaluation expérimentale

4.1. *Collection de test et protocole*

Avec l'absence d'un cadre standard d'évaluation en RIS, nous avons conçu une collection de 8433 films extraite du site 'imdb.com'. Nous avons demandé à dix utilisateurs de définir chacun une requête, et de juger la pertinence des 20 premiers documents retournés sur la base de leur pertinence thématique (*Lucene solr*) et sur la base de leur pertinence sociale. Afin d'étudier l'impact de la combinaison du score social avec le score thématique sur l'estimation globale de la pertinence, nous avons configuré les paramètres α et β à 0.2 pour maximiser le score social, ainsi que les facteurs de popularité et d'importance de la ressource par rapport à la fraicheur.

4.1. *Résultats et discussion*

Les résultats expérimentaux sont très encourageants, même s'ils ont été réalisés sur une petite collection de test. Nous présentons dans le Tableau 2 les différentes valeurs de P@20 et nDCG@20 obtenues pour l'ensemble des requêtes.

Systèmes	nDCG@20	P@20
Système 1 : *Lucene Solr*	0.76	0.38
Système 2 : *Lucene Solr* + Propriétés sociales	0.942 (+24%)	0.73

Tableau 2. *Comparaison des précisions P@20 et des moyennes du nDCG@20*

Les résultats obtenus par notre système sont meilleurs par rapport au système qui utilise uniquement la pertinence thématique. En effet, notre système améliore le nDCG à 24% par rapport à la RI classique. Ceci reflète l'efficacité d'ordonnancement social en termes de qualité des documents retournés en tête de liste. Il améliore aussi la valeur de précision de manière satisfaisante pour l'utilisateur. Ces améliorations

apportées par notre modèle montrent principalement l'intérêt de la combinaison de la pertinence thématique et la pertinence sociale, sachant que les propriétés qualitatives apportent plus de gain par rapport à la propriété temporelle.

Après cette évaluation expérimentale, nous pouvons dire que l'intégration des propriétés sociales dans un modèle de recherche améliore l'ordre de pertinence, et la précision des résultats des moteurs de recherche.

5. Conclusion

Ceci est un travail préliminaire, nous sommes tout à fait conscients que le modèle de combinaison des facteurs est très simple. Une réflexion plus poussée pour mieux répondre à ces questions est nécessaire. Il est également à noter que les expérimentations sont assez préliminaires. D'autres expérimentations à plus grande échelle sont également nécessaires. Ceci étant même avec ces éléments simples, les premiers résultats obtenus nous encouragent à investir davantage cette piste.

En perspective, nous envisageons d'intégrer d'autres propriétés sociales au modèle proposé et de faire l'évaluation sur une grande collection de test. Nous envisageons également de mettre en œuvre un processus d'apprentissage afin de définir la meilleure configuration des différents paramètres.

6. Bibliographie

Baeza-Yates R., « User Generated Content: How Good is It? », *Proceedings of the 3rd Workshop on Information Credibility On the Web WICOW'09*, *ACM*, 20 Avril 2009, Madrid Spain, p. 1-2.

Bao S., Xue G., Wu X., Yu Y., Fei B., Su Z., « Optimizing Web Search Using Social Annotations », *Proceedings of the 16th international conference on World Wide Web WWW'07*, 8-12 Mai 2007, Banff, Alberta, Canada, p. 501-510.

Ben-Jabeur L., Tamine L., Boughanem M., « Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter », *Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique MARAMI'11*, 19-21 Octobre 2011, Grenoble, France.

Hong L., Dan O., Davison B., « Predicting Popular Messages in Twitter », *Proceedings of the 20th international conference companion on World Wide Web WWW'11*, 28 Mars 2011, Hyderabad, India, p. 57-58.

Karweg B., Hütter C., Böhm K., « Evolving Social Search Based on Boukmarks and Status Messages from Social Networks », *Proceedings of the 20th ACM international Conference on Information and Knowledge Management CIKM'11*, 24-28 Octobre 2011, Glasgow, Scotland, UK, p. 1825-1834.

Pal A., Counts S., « Identifying Topical Authorities in Microblogs », *Proceedings of the fourth ACM international conference on Web Search and Data Mining WSDM'11*, 9-12 Février 2011, Hong Kong, China, p. 45-54.

Volkovich Y., Kaltenbrunner A., « Evaluation of Valuable User Generated Content on Social News Web Sites », *World Wide Web Companion Volume WWW'11*, *ACM*, 28 Avril 2011, Hyderabad, India, p. 139-140.