Exploitation de signaux sociaux pour estimer la pertinence a priori d'une ressource

Ismail Badache, Mohand Boughanem

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS, SIG
118 Route de Narbonne
F-31062 Toulouse cedex 9
France

{Ismail.Badache, Mohand.Boughanem}@irit.fr

RÉSUMÉ. *Dans cet article nous proposons une approche de recherche d'information (RI) qui prend en compte le contenu social associé à une ressource pour mesurer sa pertinence a priori vis-à-vis d'une requête. Nous démontrons comment ces caractéristiques, qui sont sous forme d'actions relevant d'activités sociales (signaux sociaux) tels que le nombre de "j'aime" et de "partage", peuvent être combinées pour quantifier des propriétés sociales telles que la popularité et la réputation. Nous proposons de modéliser ces propriétés comme des probabilités a priori que nous intégrons dans un modèle de langue. Nous avons évalué l'efficacité de notre approche sur la collection d'IMDb contenant 32706 documents et leurs caractéristiques sociales collectées sur plusieurs réseaux sociaux. Nos résultats expérimentaux sont très prometteurs et montrent l'intérêt de l'intégration des propriétés sociales dans un modèle de recherche pour améliorer la RI.*

ABSTRACT. *In this paper we propose an information retrieval (IR) approach which takes into account the social content associated with a resource to measure its a priori relevance to a query. We show how these characteristics, which are of the form of actions (social signals) such as the number of "like" and "share", can be combined to quantify social properties such as popularity and reputation. We propose to model these properties as a priori probabilities that we integrate into a language model. We evaluated the effectiveness of our approach on the IMDb dataset containing 32706 documents and their social characteristics collected from several social networks. Our experimental results are very promising and show the interest of integrating social properties in search model to enhance IR.*

MOTS-CLÉS : *Signaux sociaux, recherche d'information sociale, réseaux sociaux, propriétés sociales, modèle de tri, corrélation.*

KEYWORDS: *Social signals, social information retrieval, social networks, social properties, ranking model, correlation.*

CORIA 2014, pp. 171–186, Nancy, 18-21 mars 2014

1. Introduction

Les systèmes de recherche d'information (RI) visent à rechercher des documents pertinents en réponse à un besoin de l'utilisateur, habituellement exprimé par une requête. La plupart des modèles de recherche utilisent les statistiques des termes, telles que la fréquence du terme dans le document et dans la collection de documents. Outre ces facteurs, les modèles de RI sont souvent étendus avec d'autres sources d'évidence indépendantes de la requête qui mesurent l'importance (ou l'intérêt) *a priori* du document (Craswell *et al.,* 2005). On y trouve par le nombre de liens entrants, la longueur d'un document (Kraaij *et al.,* 2002) et le PageRank (Brin et Page, 1998).

Une des sources importantes que l'on peut également exploiter pour mesurer l'intérêt d'une page Web ou de manière générale une ressource, est le Web social. En effet, grâce aux outils proposés par le Web 2.0 les utilisateurs interagissent de plus en plus entre eux et/ou avec les ressources. Ces interactions, traduites par des annotations, des commentaires ou des votes sur des ressources, produisent de l'information sociale utile et intéressante pour caractériser une ressource, en termes de popularité, de réputation et de fraîcheur.

Dans cet article, nous proposons d'exploiter les signaux sociaux laissés par les utilisateurs sur les ressources pour mesurer la pertinence (l'intérêt) *a priori* d'une ressource. Cette connaissance *a priori* est combinée avec la pertinence thématique dans un modèle de langue qui prend en compte ces sources d'évidence. Les questions de recherche auxquelles nous souhaitons répondre dans cet article sont les suivantes :

(a) Comment traduire les signaux sociaux en propriétés sociales ?

(b) Quelles sont les propriétés sociales utiles pour évaluer la pertinence *a priori* d'une ressource?

(c) Quel modèle théorique pour combiner la pertinence *a priori* d'une ressource et sa pertinence thématique ?

(d) Impact des propriétés sociales dans les performances d'un système de recherche d'information ?

L'article est structuré de la façon suivante : nous présentons dans la section 2 un aperçu sur certains travaux connexes. Ensuite, nous détaillons notre approche sociale dans la section 3. La section 4 est consacrée à l'expérimentation effectuée sur une collection issue d'IMDb[1]. Enfin la section 5 conclut l'article et annonce des perspectives.

1. *http://www.imdb.com/*

2. État de l'art

Dans cette section, nous présentons les travaux exploitant des sources d'évidence indépendantes de la requête pour mesurer la pertinence (l'importance) *a priori* d'une ressource.

L'une des sources d'évidence largement exploitée en RI est la structure des hyperliens. Un grand nombre de liens entrants dans un document indique que de nombreux documents considèrent le document en question important ou autoritaire. Les premiers algorithmes ayant exploités cette source d'évidence sont PageRank (Brin et Page, 1998) et HITS (Kleinberg, 1999). Ils associent une valeur d'autorité à chaque page Web, cette valeur est d'autant plus forte que les documents qui pointent cette page possèdent eux aussi une valeur forte. Une autre source de connaissance *a priori* est le localisateur uniforme de ressource (URL). Chaque document sur le Web est identifié par une adresse URL qui se compose d'un nom du serveur, un chemin d'accès et un nom de fichier. (Kraaij *et al.,* 2002) ont défini quatre types d'URL. L'évaluation sur la collection TREC-Web 2001 a montré que le fait que les pages d'entrée tendent à avoir des URL plus courtes que les autres documents peut être exploité avec succès par un algorithme de classement.

Certains travaux exploitent d'autres types de caractéristiques de document issues des réseaux sociaux. (Chelaru *et al.*, 2012) étudient l'impact des signaux sociaux (*aime*, *n'aime pas*, *commentaires*, etc.) sur l'efficacité de la recherche sur "YouTube.com". Ils montrent que, bien que les critères de base basés sur la similarité de la requête avec le titre vidéo et les annotations sont efficaces pour la recherche vidéo, les critères sociaux sont également très utiles et améliorent le classement des résultats de la recherche pour 48% des requêtes. Ils ont utilisé "greedy feature selection algorithm" et six algorithmes d'apprentissage. Notre approche exploite le même principe, mais contrairement à l'approche ci-dessus, nous n'utilisons pas de techniques d'apprentissage, et nous exploitons davantage de signaux issus de réseaux sociaux multiples.

(Karweg *et al.*, 2011) proposent une approche combinant un score thématique et un score social basé sur deux facteurs : (a) premièrement, l'intensité d'engagement d'un utilisateur pendant une interaction avec un document, mesurée à partir du nombre de clics, nombre de votes, nombre d'enregistrement et recommandation; (b) deuxièmement, le degré de confiance mesuré à partir du graphe social pour chaque utilisateur selon sa popularité, en utilisant l'algorithme de *PageRank*. Ils montrent que les résultats de la recherche sociale sont plus pertinents. En particulier, ils diminuent le temps requis pour le processus de recherche et augmentent la satisfaction des utilisateurs. De même, (Khodaei et Shahabi, 2012) proposent une approche de classement fondé sur plusieurs paramètres sociaux combinés avec la pertinence textuelle classique. Ces facteurs sont l'importance des utilisateurs et des documents en fonction des relations entre les utilisateurs et les actions des utilisateurs (nombre de lectures d'un titre sur *last.fr*) effectuées sur les documents. Ils ont mené un ensemble d'expériences sur des données issues du site Internet de Radio en ligne *last.fm*. Les résultats expérimentaux ont été prometteurs et montrent

une amélioration significative pour le classement socio-textuelle par rapport au textuel.

Par rapport à la RI sur Twitter, (Alonso *et al.*, 2010) considèrent que la présence d'un lien URL est un critère important pour distinguer les *tweet* intéressants, avec plus de 80% de précision. Cependant, cette règle risque de considérer de nombreux *tweet* intéressants comme étant non intéressants, juste parce qu'ils ne contiennent pas de liens. (Yang *et al.*, 2012) considèrent qu'un *tweet* intéressant doit attirer l'attention des utilisateurs au-delà du réseau propre à l'auteur (followers) et les pousse à retweeter. De même pour (Hong *et al.*, 2011) qui qualifient le nombre de retweet comme une mesure de popularité, qui peut être exploitée au sein d'un classifieur pour prédire si de nouveaux messages seront *retweetés* à l'avenir et à quelle fréquence ? Cependant, des *tweet* banals (ex. rumeurs, sans intérêts, etc.) peuvent être très populaires tels que ceux concernant des célébrités, qui possèdent généralement un très grand nombre de followers. (Yang *et al.*, 2012) modélisent Twitter comme un graphe de nœuds utilisateur et *tweet* reliés par des liens retweet et présentent une variante de l'algorithme HITS basée sur ce graphe pour produire un classement de *tweet*. (Pal et Counts, 2011) proposent un modèle d'identification des auteurs les plus influents dans le réseau de Twitter. Cette solution, utilisée en RI, est basée sur le modèle de mélange gaussien en exploitant des données sociales à partir de Twitter, telles que : le nombre de *retweet*, nombre de *tweet* conversationnel et le nombre de *followers* actifs par rapport au sujet d'intérêt.

3. Modèle de RI sociale

Notre approche de RI consiste à exploiter les signaux sociaux comme connaissances *a priori* pour définir des propriétés sociales à prendre en compte dans un modèle de recherche. Nous nous appuyons sur un modèle de langue pour combiner la pertinence thématique de la ressource vis-à-vis de la requête et son importance, modélisée elle aussi comme une probabilité *a priori*.

3.1. *Notations*

L'information sociale que nous exploitons dans le cadre de notre modèle peut être représentée par le quintuplet < U, R, A, T, RS > où *U, R, A, T, RS* sont des ensembles finis d'instances : *Utilisateurs*, *Ressources*, *Actions*, *Temps* et *Réseaux sociaux*.

Ressources : Nous considérons une collection $R = \{D_1, D_2, \dots D_n\}$ de n ressources. Une ressource D peut être un document traditionnel comme une page Web ou une ressource Web 2.0 comme une vidéo ou toute autre entité similaire. Nous supposons qu'une ressource D peut être représentée à la fois comme par un ensemble de mots-clés textuels, soit D_m, et comme un ensemble de caractéristiques sociales réalisées sur cette ressource, $D_s = \{a_1, a_2, \dots a_m\}$.

Actions : Il existe un ensemble $A = \{a_1, a_2, \dots a_m\}$ de *m* actions (signaux sociaux) que les utilisateurs peuvent effectuer sur les ressources. Ces actions représentent la relation entre l'ensemble des utilisateurs $U = \{u_1, u_2, \dots u_l\}$ et les ressources *R*. Par exemple sur Facebook, les utilisateurs peuvent effectuer les actions relevant d'activités sociales suivantes : *publier*, *aimer*, *partager* ou *commenter*.

Temps : Il représente l'historique des actions, soit $T = \{t_{a_1}, t_{a_2}, \dots t_{a_k}\}$ l'ensemble *k* moments (date) à laquelle une action a_i a été produite. Un instant de temps *t* représente la date et l'heure (datetime) de l'action effectuée par un utilisateur *u* sur une ressource *D*.

***Réseaux sociaux* :** Il existe un ensemble $RS = \{rs_1, rs_2, \dots rs_z\}$ de *z* réseau social. Chaque réseau social spécifique contient un ou plusieurs signaux sociaux spécifiques réalisés sur une ressource *r*.

3.2. *Modèle de langue et probabilité a priori*

Nous exploitons les modèles de langues pour mesurer la pertinence d'un document vis-à-vis d'une requête. Nous utilisons un modèle classique. La probabilité qu'une ressource *D* soit pertinente étant donnée une requête *Q* est estimée de la façon suivante en recourant au théorème de Bayes :

$$Score\ (Q, D) = P(D|Q) = \frac{P(D) \cdot P(Q|D)}{P(Q)}$$

En supposant que l'ordre des documents est indépendant de P(Q) et les termes sont indépendants entre eux, on a :

$$P(D|Q) =^{rank}\ P(D) \cdot P(Q|D) = P(D) \cdot \prod_{m_i \in Q} P(m_i|D) \quad (1)$$

Où m_i représente les mots de la requêtes *Q*.

L'estimation de $P(m_i|D)$, peut être effectuée en utilisant différents modèles (Jelineck Mercer, Dirichlet). Nous avons pour notre part utilisé le modèle JM :

$$P(m_i|D) \approx\ \lambda \cdot P(m_i|D) + (1 - \lambda) \cdot P(m_i|C) \quad (2)$$

Donc :

$$P(D|Q) \approx P(D) \cdot \prod_{m_i \in Q} \lambda \cdot P(m_i|D) + (1 - \lambda) \cdot P(m_i|C) \quad (3)$$

La probabilité $P(m_i|C)$ correspond à la probabilité de tirer un mot m_i au hasard dans la collection *C*, $P(m_i|D)$ définit la probabilité de tirer un mot m_i au hasard dans le document *D* et λ est le paramètre d'interpolation. Ce dernier s'estime habituellement comme une constante. Finalement, nous devrons estimer la probabilité *a priori* de pertinence d'un document $P(D)$. Nous notons que les

modèles de langues fournissent une manière fondée théoriquement pour prendre en compte la notion de *probabilités a priori* d'un document $P(D)$.

3.3. *Estimation des probabilités a priori*

Pour estimer la probabilité *a priori* d'une ressource $P(D\ est\ Pertinent)$, nous avons étudié trois propriétés sociales : a) la popularité de la ressource; b) la réputation de la ressource et c) la fraîcheur de la ressource. Les deux premières propriétés sont quantifiées à partir du nombre d'occurrences d'une ou plusieurs actions spécifiques pour chacune des deux propriétés. Tandis que la fraîcheur est mesurée à partir des dates des dernières actions pour une ressource Web donnée.

3.3.1. *Popularité de la ressource*

La popularité d'une ressource Web $\boldsymbol{P_{Soc}}$ est un phénomène social qui dicte quel est le plus connu dans le public. Grâce à l'influence des pairs, des ressources cibles peuvent rapidement monter dans la façon dont ils sont omniprésents dans la société. Donc, la popularité peut être estimée en fonction de l'intensité de *partage* de ces ressources entre les utilisateurs à travers les signaux sociaux. Nous considérons qu'une ressource est dite populaire si elle a été *publiée* et *partagée* par plusieurs utilisateurs dans plusieurs réseaux sociaux, au point de devenir très connue dans le grand public.

3.3.2. *Réputation de la ressource*

Nous supposons que si l'indice de popularité d'une ressource est important, cette ressource est relativement intéressante. Mais la popularité d'une ressource ne reflète pas forcément sa bonne ou mauvaise réputation. La réputation $\boldsymbol{R_{Soc}}$ d'une ressource est une opinion sur cette ressource, généralement un résultat de l'évaluation sociale sur un ensemble de critères. Nous pensons que l'estimation de cette propriété peut être calculée à partir des actions relevant d'activités sociales qui portent un sens positif tel que le *j'aime* de Facebook ou le *marquage* d'une ressource comme favoris sur Delicious. En effet, la réputation d'une ressource est vue par rapport à son degré d'appréciation dans les réseaux sociaux.

Pour la suite de l'article nous utilisons les notations suivantes :

- $x \in \{P_{Soc}, R_{Soc}\}$ désigne la propriété sociale estimée à partir d'un ensemble d'actions spécifiques.
- $P_x(D) = P_x(D_s)$ représente la probabilité *a priori* relative à la propriété sociale *x*, où D_s est la ressource représentée par ses caractéristiques sociales.
- $Count(a_i^x, D_s)$ représente le nombre d'apparition d'une action spécifique a_i^x dans la ressource D_s. a_i^x désigne l'action a_i exploitée pour mesurer la propriété *x*.
- $\sum_{j=1}^{n} Count(a_i^x, D_{s_j})$ représente le nombre total d'apparition d'une action spécifiques a_i^x dans toutes les ressources retournées D_{s_j}.

- $Count(a_i^x, C)$ représente le nombre d'apparition d'une action spécifique a_i^x dans la collection *C*.
- $\sum_{i=1}^{m} Count(a_i^x, C)$ représente le nombre total d'apparition de toutes les actions spécifiques a_i^x dans la collection *C*.

Nous définissons la probabilité *a priori* relative à la popularité/réputation d'une ressource par la formule suivante :

$$P_x(D) = P_x(D_s) = \prod_{a_i^x \in A} P_x(a_i^x) = \prod_{a_i^x \in A} c \cdot Count(a_i^x, D_s) \quad (4)$$

c est une constante permettant de traduire le simple comptage la fonction *Count()* en distribution de probabilité. Cette constante peut être relative au document ou à une collection de documents ou encore à un type de signal. Pour nombre part, dans cet article, nous considérons cette constante relativement au sous-ensemble de documents renvoyés par une requête, elle est traduite par $\sum_{j=1}^{n} Count(a_i^x, D_{s_j})$. Elle est de ce fait dépendante juste du type de signal considéré. En appliquant un lissage Dirichlet (Zhai et Lafferty, 2004) on aura :

$$P_x(D_s) = \prod_{a_i^x \in A} \frac{Count(a_i^x, D_s) + \mu \cdot P(a_i^x|C)}{\sum_{j=1}^{n} Count(a_i^x, D_{s_j}) + \mu} \quad (5)$$

Où $P(a_i^x|C)$ estime le nombre de fois qu'une action spécifique a_i^x apparait dans la collection *C* relativement à l'ensemble des actions de la collection. μ est un paramètre appelé pseudo-fréquence. La formule est comme suit :

$$P(a_i^x|C) = \frac{Count(a_i^x, C)}{\sum_{i=1}^{m} Count(a_i^x, C)} \quad (6)$$

Exemple : soient deux ressources D_1 *et* D_2 représentées par les signaux sociaux suivants :

	J'aime	+1	Partage	Commentaire
D_1	155	78	49	250
D_2	100	88	310	120
$\sum_{j=1}^{m} Count(a_i^x, C)$	50000	20000	60000	30000
$\sum_{j=1}^{n} Count(a_i^x, D_{s_j})$	1300	1100	1500	1200

Nous estimons dans cet exemple la popularité P_{Soc} par rapport aux nombres de *Commentaire* et de *Partage*, tant dis que la réputation R_{Soc} est estimée par rapport

aux nombres de +1 et de *J'aime*. En appliquant la formule 5 pour les deux propriétés sociales on aura :

- *Document D_1* :

$$P_{P_{Soc}}(D_{s_1}) = \frac{49 + 250 \cdot \frac{60000}{60000 + 30000}}{1500 + 250} \cdot \frac{250 + 250 \cdot \frac{30000}{30000 + 60000}}{1200 + 250} = 0.0283$$

$$P_{R_{Soc}}(D_{s_1}) = \frac{155 + 250 \cdot \frac{50000}{50000 + 20000}}{1300 + 250} \cdot \frac{78 + 250 \cdot \frac{20000}{20000 + 50000}}{1100 + 250} = 0.0238$$

- *Document D_2* :

$$P_{P_{Soc}}(D_{s_2}) = \frac{310 + 250 \cdot \frac{60000}{60000 + 30000}}{1500 + 250} \cdot \frac{120 + 250 \cdot \frac{30000}{30000 + 60000}}{1200 + 250} = 0.0381$$

$$P_{R_{Soc}}(D_{s_2}) = \frac{100 + 250 \cdot \frac{50000}{50000 + 20000}}{1300 + 250} \cdot \frac{88 + 250 \cdot \frac{20000}{20000 + 50000}}{1100 + 250} = 0.0212$$

3.3.3. *Fraîcheur de la ressource*

La fraîcheur d'une ressource $\boldsymbol{F_{Soc}}$ est un facteur de pertinence important, exploité par plusieurs moteurs de recherche. La fraîcheur d'une information revient souvent à sa date de publication, mais nous ne pouvons pas dire qu'une information est obligatoirement obsolète parce qu'elle a été publiée il y a deux ans. Nous supposons qu'une ressource est dite fraiche, si des signaux sociaux très récents ont été associés à cette ressource. Pour cela, nous proposons cette définition de *fraîcheur*: « *La date de la dernière interaction (ex. commentaire, mention, etc.) avec une ressource dans les réseaux sociaux, peut être utilisée pour mesurer la fraîcheur de l'information* ».

$$P_{F_{Soc}}(D) = P_{F_{Soc}}(D_s) = \prod_{a_i \in A} P_{F_{Soc}}(t_{a_i}) = \prod_{a_i \in A} \frac{1}{Time(t_{a_i}, D_s)} \quad (7)$$

Où :

- t_{a_i} représente le moment (date) à laquelle la dernière action a_i du même type a été produite.
- $P_{F_{Soc}}(t_{a_i})$ représente la récence d'une action a_i dans la ressource D_s, d'où nous estimons la fraîcheur de la ressource en elle-même.
- $Time(t_{a_i}, D_s) = t_{Actuel} - t_{a_i}$ estime le temps écoulé depuis la dernière action a_i du même type pour une ressource D_s. Nous notons que pour chaque action, la date t_{a_i} est initialisée par défaut à : 01-01-1970 00:00:00.

Exemple : soient $t_{Partage}$ = 2013/05/23 00:00:00 et $t_{Comment}$ = 2013/05/22 15:00:00 pour D_1 et $t_{Partage}$ = 2013/05/24 11:00:00, $t_{Comment}$ = 2013/05/25 19:00:00 pour D_2, en calculant le temps écoulé par rapport à la date actuelle (ex. 22/12/2013 00:00:00) nous aurons :

	$Time(t_{partage}, D_s)$	$Time(t_{Comment}, D_s)$
D_1	24 heures	15 heures
D_2	59 heures	91 heures
D_3	80 heures	-aucun commentaire-

- *Document D_1* :

$$P_{F_{Soc}}(D_{s_1}) = \frac{1}{24} \cdot \frac{1}{15} = 0.002$$

- *Document D_2* :

$$P_{F_{Soc}}(D_{s_2}) = \frac{1}{59} \cdot \frac{1}{91} = 1.86 \times 10^{-4}$$

- *Document D_3* : $Time(t_{Comment}, D_{s_3}) = 385463\ heures$ donc :

$$P_{F_{Soc}}(D_{s_2}) = \frac{1}{80} \cdot \frac{1}{385463} = 3.24 \times 10^{-8}$$

3.4. *Combinaison des probabilités a priori*

Dans notre cas, nous disposons de diverses sources d'informations sociales qui influence la probabilité *a priori* de pertinence. Cette probabilité est calculée par la combinaison de plusieurs propriétés sociales (*Fraîcheur*, *Popularité* et *Réputation*), dont ces dernières sont quantifiées par des caractéristiques sociales. De manière générale, le problème peut être formalisé comme suit (Peng *et al.*, 2007) :

$$P_{F_{Soc} \oplus P_{Soc} \oplus R_{Soc}}(D_s) = P_{F_{Soc}}(D_s) \cdot P_{P_{Soc}}(D_s) \cdot P_{R_{Soc}}(D_s) \quad (8)$$

Où $P_{F_{Soc}}(D_s)$, $P_{P_{Soc}}(D_s)$ et $P_{R_{Soc}}(D_s)$ sont les probabilités *a priori* d'une ressource relative à F_{Soc}, P_{Soc} et R_{Soc}, respectivement. $P_{F_{Soc} \oplus P_{Soc} \oplus R_{Soc}}(D_s)$ est la probabilité de la combinaison des trois probabilités *a priori*.

4. Evaluation expérimentale

Afin de valider notre modèle, nous avons effectué une série d'expérimentations sur la collection IMDb (Internet Movie Database). En effet, nous faisons une étude sur l'apport des propriétés sociales au modèle de recherche basé uniquement sur le contenu. Dans nos expérimentations nous avons utilisé le modèle de RI décrit dans la section 3.2. Les probabilités *a priori* dans notre modèle soit $P(D_s)$ sont définies dans la section 3.3, mais quantifiées de la façon présentée dans la section 4.2. Nous détaillons dans ce qui suit notre protocole d'expérimentation.

4.1. *Description de la collection de test*

Nous avons collecté 32706 documents en anglais extraits du site "imdb.com". Chaque document décrit un film, et est représenté par un ensemble de métadonnées, listées dans le Tableau 1. Chaque document a été indexé en fonction des mots clés se trouvant dans les balises ayant le statut indexé dans le Tableau 1. L'indexation est classique, utilisation de *Porter* et les mots vides sont supprimés.

Champ	Description	Statut
ID	identifiant du film (le document).	-
Title	le titre du film.	indexé
Year	l'année de sortie du film.	indexé
Rated	classement des films selon le type du contenu.	-
Released	date de réalisation du film.	indexé
Runtime	durée du film.	indexé
Genre	genre de film (Action, Drame, etc.).	indexé
Director	le directeur du projet du film.	indexé
Writer	les écrivains et les scénaristes du film.	indexé
Actors	les acteurs principaux du film.	indexé
Plot	résumé textuel du film.	indexé
Poster	le lien URL de l'affiche du film.	-
url	le lien URL qui mène à la source originale du document.	-
UGC	Les différents signaux sociaux récupérés.	-

Tableau 1. *Liste des différents champs d'un document dans la collection*

A chaque document est également associé un certain nombre de signaux sociaux. Nous l'avons mis dans la balise UGC (User Generated Content). Ce champ n'a pas été indexé. Le tableau 2 donne quelques statistiques sur le nombre de signaux sociaux dans la collection de documents.

Network	Signaux sociaux	Somme	Min	Max	Moyenne
Facebook	*J'aime*	5056517	0	79693	154
	Partage	5778414	0	41618	176
	Commentaire	6717573	0	60081	205
Twitter	*Tweet*	1097204	0	22954	33
Google+	+1	139189	0	1368	4
Delicious	*Marque*	32810	0	1033	1
LinkedIn	*Partage*	57545	0	25215	1

Tableau 2. *Statistiques sur le nombre des signaux sociaux dans la collection*

4.2. *Quantification des propriétés sociales*

Afin de quantifier ces propriétés sociales, nous associons les signaux sociaux pour chaque propriété sociale comme suit (Tableau 3).

Propriétés sociales	Signaux sociaux	Réseaux sociaux
Popularité (P_{Soc})	Nombre de « *Commentaire* »	Facebook
	Nombre de « *Tweet* »	Twitter
	Nombre de « *Partage* »	LinkedIn, Facebook
Réputation (R_{Soc})	Nombre de « +1 »	Google+
	Nombre de « *J'aime* »	Facebook
	Nombre de « *Marque* »	Delicious
Fraîcheur (F_{Soc})	Date de la dernière *action*	Facebook

Tableau 3. *Liste des signaux sociaux exploités dans la quantification*

Les signaux sociaux ont été associés pour chaque propriété selon leur nature et signification, qui correspondent à la définition que nous avons donnée en section 3.2. Dans le tableau 3, nous remarquons que les signaux sociaux estimant la réputation portent des opinions positives, par exemple, *marquer* un lien d'une ressource par un utilisateur sur Delicious signifie que ce lien a été rajouté à sa liste de favoris. Pour le *j'aime* et *+1*, l'utilisateur clique sur ces deux boutons pour indiquer qu'il a apprécié le contenu de cette ressource. Donc la présence de cet ensemble de signaux sociaux dans une ressource augmente le degré de réputation de cette ressource. De même pour la popularité, les signaux sociaux exploités pour estimer cette dernière, nous permettent de savoir la position en termes de tendance et propagation de cette ressource sur le Web. Enfin, comme les dates des derniers signaux sociaux effectués ne sont pas disponibles, l'estimation de la fraîcheur est calculée à partir des dates des derniers *partages* et *commentaires* sur Facebook.

Le tableau 4 montre un exemple de signaux sociaux pour quelques documents. L'URL du document est donnée par la syntaxe suivante: www.imdb.com/title/{id}/

Réseaux	**Facebook**			**Google+**	**Twitter**	**Delicious**	**LinkedIn**
Id	*J'aime*	*Partage*	*Commentaire*	+1	*Tweet*	*Marque*	*Partage*
tt1730728	30	11	2	0	0	0	0
tt1922777	12363	11481	20614	238	2522	12	14
tt1925050	0	2	7	0	0	0	0

Tableau 4. *Statistique sur le nombre des signaux sociaux de 3 documents*

4.3. *Requêtes et jugement de pertinence*

Nous avons défini 25 requêtes, dont 20 requêtes sont issues de la collection INEX IMDb et 5 sont créés par nous-même. Le tableau suivant montre un exemple de requêtes:

Requête	Description
"true story event movies"	recherche des films basés sur des évènements inspirés d'une histoire vraie.
"christmas familly movies"	recherche des films familiaux de noël.
"Martial arts sport movies documentary"	recherche de documentaires ou de films sur les arts martiaux ou toutes activités sportives.

Tableau 5. *Exemple de requêtes d'évaluation*

Pour obtenir les jugements de pertinence des documents, nous avons impliqué 12 participants. Il y avait 6 femmes et 6 hommes, la tranche d'âge était entre 23 et 31 ans. Tous les participants étaient de la discipline informatique dont 4 sont des étudiants, 3 doctorant, 2 enseignants, et le reste sont des développeurs.

Nous avons demandé à chaque participant de choisir trois requêtes de notre ensemble de requêtes. Ensuite, nous avons attribué à chaque participant deux ou trois requêtes et leur avons demandé d'évaluer les 100 premiers documents affichés pour une requête donnée en utilisant une échelle d'évaluation à 3 points (non-pertinent, peu pertinent et pertinent). Nous notons que chaque requête est jugée par 3 utilisateurs. Pour éviter tout biais, aucun des signaux sociaux n'a été affiché avec les documents, mais toutes les métadonnées textuelles sont affichées pour faciliter la tâche de jugement.

Pour évaluer la performance de notre approche, nous calculons les métriques de précision, le nDCG (Jarvelin et Kekäläinen, 2002) et la MAP. Néanmoins, il est intéressant d'examiner combien d'accord entre les juges y a eu sur les jugements de pertinence. Dans les sciences sociales, une mesure commune d'un accord entre les juges représente la statistique de kappa (Cohen, 1960).

4.4. *Résultats et discussion*

Nous comparons notre approche avec deux modèles de base, exploitant uniquement le contenu du document, soient BM25 (Robertson et Walker, 1994) et le modèle de langue *Hiemstra* (1998). Nous avons utilisé le modèle de Hiemstra, comme modèle de base pour le calcul de la pertinence thématique. En ce qui concerne le paramétrage, nous avons attribué la valeur par défaut 0.85 pour le paramètre λ (voir formule 3). La valeur optimale du paramètre μ semble varier d'une collection à l'autre, bien que dans la plupart des cas elle est autour de 2000 (Zhai et Lafferty, 2004), dans notre cas nous avons attribué la valeur 250 pour le paramètre de lissage μ. Les résultats obtenus sont présentés dans le tableau suivant :

Métriques / Modèles RI	P@10	nDCG@10	P@20	nDCG@20	MAP
BM25	0.5101	0.5948	0.4511	0.5748	0.2601
ML-Hiemstra	0.5487	0.6167	0.4754	0.5866	0.2666
J'aime	0.6211*	0.7066*	0.5336	0.6414	0.3155*
Partage	0.6135	0.6842	0.5371*	0.6586*	0.3126*
Commentaire	0.6100*	0.6637*	0.5188*	0.6507*	0.2912*
Tweet	0.6018	0.6354	0.5077	0.6048	0.2844
+1	0.5563	0.6189	0.4857	0.5979	0.2670
Marque	0.5704	0.6207	0.4787	0.5965	0.2697
Partage (LinkedIn)	0.5800	0.6178	0.4913	0.5921	0.2701
ML+*Popularité* (P_{Soc})	0.6800*	0.7471*	0.5925*	0.6951*	0.3333*
ML+*Réputation* (R_{Soc})	0.6813*	0.7214*	0.5819*	0.6661*	0.3219*
ML+*Fraîcheur* (F_{Soc})	0.6044	0.6911	0.5222	0.6229	0.2902
ML+(P_{Soc})+(R_{Soc})+(F_{Soc})	**0.7568**	**0.7822**	**0.6577**	**0.7438**	**0.3882**

Tableau 6. *Comparaison des résultats P@{10, 20}, nDCG@{10,20} et MAP*

Le Tableau 6 récapitule les résultats de précision et de nDCG (*Normalized Discounted Cumulative Gain*) pour les 10 et 20 premiers documents retournés, ainsi que la MAP (*Mean Average Precision*). Nous remarquons dans tous les cas, avec la prise en compte des caractéristiques sociales, les résultats obtenus sont significativement meilleurs que ceux obtenus par les modèles de base. Nous constatons que la prise en compte de chaque signal social individuellement améliore les résultats, mais ils sont moins bons que ceux obtenus par la combinaison partielle/globale de ces caractéristiques sociales. Nous avons remarqué aussi que la mesure de l'accord Kappa (Cohen, 1960) varie entre 0.69 et 0.86 pour l'ensemble des requêtes. La valeur moyenne de l'accord entre les évaluateurs est de 81.24%, ce qui correspond à un bon accord.

Nous constatons également que les meilleurs résultats sont obtenus avec la prise en compte des signaux sociaux combinés sous forme de propriétés qualitatives (*popularité* et *réputation*) et temporelle. Enfin la combinaison des trois propriétés est la configuration qui apporte les meilleurs résultats comparativement à toutes les autres configurations. La *fraîcheur* dans notre étude est vue par rapport à la récence des actions relevant d'activités sociales, le nombre d'action sur une ressource est lié à sa *fraîcheur* dans les réseaux sociaux, mais des ressources qui possèdent des signaux plus frais sont supposées être mieux classées. Enfin, on constate également que la prise en compte de la *fraîcheur* seule et combinée avec les autres signaux améliore également les résultats. Il est par ailleurs à noter que la *fraîcheur* dans notre cas est corrélée à la présence de signaux sociaux. Donc une question demeure posée, est-ce juste la présence du signal qui améliore la pertinence ou bien c'est la

fraîcheur du signal qui contribue également à cette amélioration. Des analyses un peu plus approfondies sont nécessaires pour répondre cette question.

4.5. *Analyse de corrélation des rangs*

Selon une étude en Juin 2013 par *Searchmetrics*[2], les signaux sociaux représentent 5 des 6 facteurs les plus fortement corrélés avec les résultats de recherche Google. En outre, l'enquête *BrightEdge*[3] publié en Janvier 2012, révèle que 84% des marketeurs de recherche disent que les signaux sociaux tels que le *j'aime*, *tweet*, et *Google*+1 seront soit plus important (53%) ou beaucoup plus important (31%) à leur référencement cette année par rapport à 2011 et 2012.

Nous avons effectué une analyse de corrélation des rangs en utilisant le coefficient de corrélation de *Spearman* (Bolboaca et Jäntschi 2006) entre les signaux sociaux des documents et le degré de pertinence. La figure 1 présente les valeurs de corrélations des rangs des signaux sociaux par rapport au degré de pertinence des documents. L'étude montre que le *j'aime* (0.29) a la plus forte corrélation, suivi par le nombre de *commentaire* (0.28). D'autres caractéristiques de haut rang incluent le nombre de *partage* (0.27) et le nombre de *tweet* (0.23).

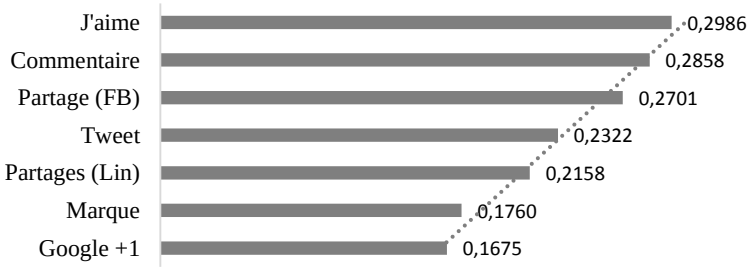

Figure 1. *Corrélation des signaux sociaux*

La figure 2 démontre que la combinaison des signaux sociaux selon leur nature pour définir les propriétés sociales (*popularité*, *réputation* et *fraîcheur*) augmente le taux de corrélation des rangs par rapport à la pertinence des documents.

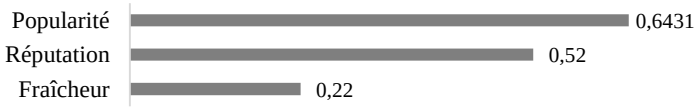

Figure 2. *Corrélation des propriétés sociales*

2. *www.searchmetrics.com/en/services/ranking-factors-2013/*
3. *www.marketingcharts.com/direct/social-signals-increasingly-important- to-seo-20695/*

L'analyse de corrélation des rangs nous montre que tous les signaux sociaux présentent une corrélation positive. Notre étude confirme l'intérêt d'exploiter les signaux sociaux.

5. Conclusion

Nous avons proposé dans cet article un modèle de recherche des ressources Web basé sur des propriétés sociales. Ces propriétés, considérées comme des probabilités *a priori*, ont été définies à partir des signaux sociaux. Le modèle proposé s'appuie sur un modèle de langue qui incorpore ces nouvelles connaissances *a priori*. L'évaluation expérimentale menée sur la collection IMDb montre que la prise en compte de ces propriétés sociales au sein d'un modèle de recherche textuel permet d'améliorer la qualité des résultats de recherche retournés.

En perspective, nous prévoyons de répondre à certaines limitations de l'étude en cours. Tout d'abord, nous envisageons d'estimer d'autres types de connaissances *a priori* et d'étudier la distribution de ces contenus sociaux dans les documents Web. Ensuite, nous comptons améliorer la manière de prendre en compte le temps dans le calcul de l'intérêt de la ressource. D'autres expérimentations à plus grande échelle sur d'autres types de collections sont également envisagées. Ceci étant même avec ces éléments simples, les premiers résultats obtenus nous encouragent à investir davantage cette piste.

6. Bibliographie

Alonso O., Carson C., Gerster D., Ji X., Nabar S., « Detecting Uninteresting Content in Text Streams », *SIGIR*, 19-23/07/2010, Switzerland, p. 17-22.

Alonso O., Gamon M., Haas K., Pantel P., « Diversity and Relevance in Social Search», *DDR*, 12/02/2011, Seattle USA, p. 1-3.

Badache I., «RI sociale: intégration de propriété sociale dans un modèle de recherché» , *CORIA*, 2013, Neuchâtel, Switzerland, p. 463-468.

Bolboaca S.D., Jäntschi L., « Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds », *Leonardo Journal of Sciences*, 2006, vol. 5, no 9, p. 179-200.

Brin S., Page L., « The anatomy of a large-scale hypertextual Web search engine », *Computer Networks and ISDN Systems*, vol. 30, n°1–7, 1998, p. 107–117.

Chelaru S., Orellana C., Altingovde I., « Can Social Features Help Learning to Rank Youtube Videos? », *WISE'12*, 28-30/11/2012, Paphos, Cyprus, p. 552-566.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and Psychological Measurement*, Vol 20 (1), 1960, p. 213–220.

Craswell N., Robertson S., Zaragoza H., Taylor M., « Relevance weighting for query independent evidence », *SIGIR*, 2005, p. 416-423.

Evans B.M., Chi E.H., « Towards a Model of Understanding Social Search », *CSCW, ACM*, 08-12/11/2008, California, USA, p. 485-494.

Järvelin K., Kekäläinen J., « Cumulated gain-based evaluation of information retrieval techniques », *ACM Transactions on Information Systems*, Vol 20 (4), 2002, p. 422–446.

Hawking D., Craswell N., « Overview of the TREC-2001 Web Track », *TREC*, 2001, p. 25–31, 2001.

Hiemstra D., « A linguistically motivated probabilistic model of information retrieval», *ECDL*, Septembre 1998, Springer Verlag New York USA.

Hong L., Dan O., Davison B., « Predicting Popular Messages in Twitter », *WWW*, 28/03/2011, India, p. 57-58.

Horowitz D., Kamvar S.O., « The Anatomy of a Large-scale Social Engine », *WWW*, *ACM*, 26-30/04/2010, North California, USA, p. 431-440.

Karweg B., Hütter C., Böhm K., « Evolving Social Search Based on Boukmarks and Status Messages from Social Networks », *CIKM*, 24-28/10/2011, Scotland, UK, p. 1825-1834.

Khodaei A., Shahabi C., « Social-Textual Search and Ranking », *CrowdSearch Workshop at WWW*, 2012, Lyon, France, p. 3-8.

Kleinberg J. M., « Authoritative sources in a hyperlinked environment », *Journal of the ACM*, vol. 46, n° 5, p. 604–632, 1999.

Kraaij W., Westerveld T., Hiemstra F., « The importance of prior probabilities for entry page search ». *SIGIR*, 2002, p. 27-34.

Pal A., Counts S., « Identifying Topical Authorities in Microblogs », *WSDM*, 9-12/02/2011, China, p. 45-54.

Peng J., Macdonald C., He B., Ounis I., « Combination of document priors in Web information retrieval », RIAO, 2007., Paris, France, p. 596-611.

Robertson S.E., Walker S., « Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval», *SIGIR*, USA. Springer-Verlag New York, p. 232–241.

Yang M.C., Lee JT., Lee S.W., Rim H.C., « Finding Interesting Posts in Twitter Based on Retweet Graph Analysis », *SIGIR*, 12-16/08/2012, Portland.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to information retrieval », *ACM Trans. Inf. Syst.* Vol 22 (2), April 2004, p. 179-214.