HAL
open science

Pertinence a Priori Basée sur la Diversité et la Temporalité des Signaux Sociaux

Ismail Badache, Mohand Boughanem

To cite this version:

Ismail Badache, Mohand Boughanem. Pertinence a Priori Basée sur la Diversité et la Temporalité des Signaux Sociaux. Conférence francophone en Recherche d'Information et Applications (CORIA 2015), ARIA (Association Francophone de Recherche d'Information (RI) et Applications), Mar 2015, Paris, France. pp.23–38. hal-01276255

HAL Id: hal-01276255

https://hal.science/hal-01276255v1

Submitted on 19 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 14817

To cite this version : Badache, Ismail and Boughanem, Mohand *Pertinence a Priori Basée sur la Diversité et la Temporalité des Signaux Sociaux.* (2015) In: Conférence francophone en Recherche d'Information et Applications (CORIA), 18 March 2015 - 20 March 2015 (Paris, France).

Any correspondance concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

Pertinence a Priori Basée sur la Diversité et la Temporalité des Signaux Sociaux

Ismail Badache et 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. *Les signaux sociaux associés aux ressources web peuvent être considérés comme une information additionnelle qui peut jouer un rôle pour mesurer une importance a priori de la ressource indépendamment de la requête. Dans cet article, nous nous intéressons particulièrement à la temporalité associée à ces signaux ainsi qu'à leur diversité. Nous supposons que l'importance a priori d'un document (ressource) dépend non seulement de la qualité de ces signaux mais aussi de la date de leur création, leur diversité ainsi que la date de publication de la ressource. De ce fait, plutôt que d'estimer cette importance (probabilité) a priori par un simple comptage des signaux liés au document, nous intégrons également la date de publication de la ressource, pour ne pas pénaliser les nouvelles ressources, les dates des signaux pour privilégier les signaux récents, ainsi que la diversité de ces signaux. Nous évaluons la performance de notre approche sur la collection d'IMDb contenant 167438 ressources et leurs données sociales collectées à partir de plusieurs réseaux sociaux. Nos résultats montrent l'intérêt des signaux temporellement sensibilisés à la sélection des ressources pertinentes.*

ABSTRACT. *Social signals associated with web resources can be considered as an additional information that can play a role to measure a priori importance of the resource regardless of the query. In this paper, we are particularly interested in the temporality associated with these signals and their diversity. We assume that the a priori importance of a document (resource) depends not only on the quality of these signals, but also on the dates of their creation, their diversity and the publication date of the resource. Therefore, rather than estimating the significance (probability) a priori by simply counting the number of signals associated to a resource, we also integrate the publication date of the resource, to avoid penalizing recent resources, the date of signals to boost recent actions, as well as their diversity. We evaluate the effectiveness of our approach on IMDb dataset containing 167438 resources and their social data collected from social networks. Our experiments show the interest of temporally-aware signals at capturing relevant resources.*

MOTS-CLÉS : *Signaux sociaux, Date du signal, Date de Publication de la ressource, Diversité.*

KEYWORDS: *Social signals, Signal date, Resource publication date, Diversity.*

1. Introduction

Les systèmes de recherche d'information exploitent dans leur majorité deux classes de sources d'évidences pour trier les documents répondant à une requête. La première, la plus exploitée, est dépendante de la requête, elle concerne toutes les caractéristiques relatives à la distribution des termes de la requête dans le document et dans la collection (tf-idf). La seconde classe concerne des facteurs indépendants de la requête, elle mesure une sorte de qualité ou d'importance a priori du document. Parmi ces facteurs, on distingue le PageRank (Page *et al.*, 1998), la localité thématique du document (Davison, 2000), la présence d'URL dans le document (Westerveld *et al.*, 2002), ses auteurs (Macdonald et Ounis, 2006) et les signaux sociaux (+1, j'aime, etc.) associés (Badache et Boughanem, 2014b).

En particulier, les signaux sociaux sont de plus en plus exploités par les moteurs de recherche (Sullivan, 2010). En effet, un *J'aime* ou un *+1* effectué sur une ressource peuvent être vus comme un vote, un signal, qui distingue la ressource concernée vis-à-vis d'une autre. De plus, plus ces signaux sont fréquents sur une ressource plus son importance a priori croît (Badache et Boughanem, 2014b). Cependant, dans les travaux existants les signaux sociaux sont pris en compte indépendamment du moment où l'action (le signal) s'est produite et de la date de publication de la ressource. Ils sont pris en compte uniquement par rapport à leur fréquence dans la ressource.

Pour notre part, nous supposons que l'impact d'un signal social dépend également du temps, c'est-à-dire la date à laquelle l'action de l'utilisateur est réalisée. Nous considérons que les signaux récents devraient avoir un impact supérieur vis-à-vis des signaux anciens dans le calcul de l'importance d'une ressource. La récence des signaux peut indiquer certains intérêts récents à la ressource. Ensuite, nous considérons que le nombre de signaux d'une ressource doit être pris en compte au regard de l'âge (date de publication) de cette ressource. En général, une ressource ancienne en termes de durée d'existence a de fortes chances d'avoir beaucoup plus de signaux qu'une ressource récente. Ceci conduit donc à pénaliser les ressources récentes vis-à-vis de celles qui sont anciennes. Enfin, nous proposons également de prendre en compte la diversité des signaux sociaux au sein d'une ressource.

Nous proposons dans cet article un modèle d'estimation de l'importance a priori d'une ressource qui tient compte la diversité et les caractéristiques temporelles des actions des utilisateurs comme connaissance a priori dans un modèle de recherche. Nous exploitons pour notre part un modèle de langue pour évaluer la pertinence d'un document vis-à-vis d'une requête. Ainsi, au lieu de supposer l'uniformité des probabilités a priori des documents dans ce modèle, nous attribuons des probabilités a priori estimées en fonction de la diversité de ces signaux biaisés par leur date de création et la date de publication de la ressource. Les questions de recherche abordées dans cet article sont les suivantes :

1) Comment prendre en compte les signaux sociaux et leur date de création pour estimer la probabilité a priori de la ressource ?

2) Comment estimer la diversité des signaux sociaux au sein d'une ressource ?

3) Quel est l'impact de la diversité et du temps associés aux signaux sociaux sur la performance de la RI ?

Le reste de cet article est organisé comme suit : nous présentons dans la section 2 certains travaux connexes. Ensuite, nous détaillons notre approche dans la section 3. Dans la section 4, nous évaluons l'efficacité de notre approche et nous discutons les résultats. Enfin, nous concluons ce papier en annonçant des perspectives.

2. État de l'art

Plusieurs travaux ont été proposés pour prendre en compte les signaux comme source additionnelle pour évaluer la pertinence d'un document vis-à-vis d'une requête.

Les principaux travaux existants tels que (Kazai et Milic-Frayling, 2009), (Karweg *et al.*, 2011), (Chelaru *et al.*, 2012), (Khodaei et Shahabi, 2012), (Badache et Boughanem, 2014a) et (Badache et Boughanem, 2014b) se focalisent principalement sur la façon d'améliorer l'efficacité de la RI en exploitant les actions des utilisateurs et de leurs réseaux sociaux indépendamment du temps. Par exemple, dans (Badache et Boughanem, 2014a) et (Badache et Boughanem, 2014b) nous avons montré l'impact des différents signaux sociaux pris en compte individuellement et groupés sous forme de propriétés sociales sans prise en compte de l'aspect temporel ou leur diversité dans le document.

Les travaux les plus liés à notre approche correspondent à ceux réalisés par (Inagaki *et al.*, 2010) et (Khodaei et Alonso, 2012), qui exploitent la dimension temporelle dans le classement des résultats de recherche. (Inagaki *et al.*, 2010) proposent d'exploiter la caractéristique de clic temporelle, appelé *ClickBuzz*, qui capte l'intérêt que suscite un document à travers le temps. Cette méthode permet d'exploiter le *feedback* des utilisateurs pour améliorer le processus d'apprentissage de l'ordonnancement par récence en favorisant les URL qui ont un intérêt récent pour les utilisateurs. L'utilisation de *ClickBuzz* dans les modèles de classement améliore principalement le nDCG. (Khodaei et Alonso, 2012) considèrent que la grande masse des contenus générés par les utilisateurs dans les réseaux sociaux offre une occasion d'examiner comment les utilisateurs produisent et consomment ce type de contenu au fil du temps. Ils classent les intérêts sociaux des utilisateurs en cinq classes : "recent", "ongoing", "seasonal", "past" et "random", puis analysent Twitter ainsi que des données de Facebook sur les activités sociales des usagers. Ils discutent également trois solutions différentes où ces signaux sensibles au temps peuvent être appliqués : a) la RI personnalisée ; b) la RI basée sur les amis et c) la RI collective.

Nos travaux se différencient de l'état de l'art. Notre approche est basée sur la temporalité des signaux et leur diversité dans un document. Nous notons que dans les travaux précédents la diversité a été appliquée uniquement au contenu thématique du document (Angel et Koudas, 2011) (Qin *et al.*, 2012). Dans cet article, notre but est d'estimer l'importance d'une ressource en prenant en compte sa date de publication, la récence des signaux sociaux associés ainsi que leur diversité au sein de cette ressource.

Ces sources d'évidences sont incorporées dans un modèle de langue qui fournit une solution fondée théoriquement pour prendre en compte la notion des probabilités a priori dans l'estimation de la pertinence d'un document.

3. Diversité et temporalité des signaux sociaux

Nous proposons d'estimer l'importance sociale d'une ressource en exploitant la diversité des signaux sociaux associés et le moment où l'interaction (signal) s'est produite. Afin de prendre en compte cette importance dans l'évaluation de pertinence, nous nous appuyons sur les modèles de langue pour combiner la pertinence textuelle d'une ressource vis-à-vis d'une requête et son importance socio-temporelle modélisée comme une probabilité a priori.

3.1. *Préliminaires et contexte*

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*.

3.1.1. *Ressources*

Nous considérons une collection $R = \{D_1, D_2, ...D_n\}$ 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 un ensemble de mots-clés textuels, soit $D_w = \{w_1, w_2, \ldots w_z\}$, et comme un ensemble de caractéristiques sociales réalisées sur cette ressource, $D_a = \{a_1, a_2, \ldots a_m\}$.

3.1.2. *Actions*

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

3.1.3. *Temps*

Le temps T intervient à deux niveaux dans notre approche. Il représente deux types de dimensions temporelles :

1) L'historique de chaque action, soit $T_{a_i} = \{t_{1,a_i}, t_{2,a_i}, ...t_{k,a_i}\}$ l'ensemble de k moments (date) à laquelle une action a_i a été produite. Un instant de temps t_{k,a_i} représente la date et l'heure (datetime) de l'action effectuée par un utilisateur u sur une ressource D.

2) La date de publication de la ressource, soit $T_D = \{t_{D_1}, t_{D_2}, ...t_{D_n}\}$ l'ensemble de n date à laquelle chaque ressource D de la collection R a été créée. t_D est la date de publication de la ressource D, cette date est fournie en format datetime.

3.2. *Modèle de langue et probabilités a priori*

Nous exploitons les modèles de langues (Ponte et Croft, 1998) 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 par rapport une requête Q est estimée de la façon suivante :

$$P(D|Q) \overset{\text{rank}}{=} P(D) \cdot P(Q|D) = P(D) \cdot \prod_{w_i \in Q} P(w_i|D) \qquad [1]$$

Où w_i représente les mots de la requêtes Q.

$P(D)$ représente la probabilité a priori du document D, son utilité est de modéliser et intégrer d'autres sources d'évidence indépendantes de la requête dans le processus de la recherche d'information. L'estimation de $P(w_i|D)$ peut être effectuée en utilisant différents modèles (ex. Jelineck Mercer, Dirichlet) (Zhai et Lafferty, 2004). Finalement, la principale contribution de cet article est sur l'estimation de $P(D)$ en exploitant les signaux sociaux et leurs caractéristiques temporelles.

3.3. *Estimation des probabilités a priori*

Une manière simple d'estimer la probabilité a priori est d'effectuer un simple comptage du nombre d'actions spécifiques effectuées sur une ressource. En supposant que les actions sont indépendantes entre elles, la formule générale est la suivante :

$$P(D) = \prod_{a_i \in A} P(a_i) \qquad [2]$$

$P(a_i)$ est estimée en utilisant le maximum de vraisemblance :

$$P(a_i) = \frac{Count(a_i, D)}{Count(a_\bullet, D)} \qquad [3]$$

Pour éviter une probabilité nulle, nous lissons $P(a_i)$ par la collection R en utilisant Dirichlet (Zhai et Lafferty, 2004). La formule devient comme suit :

$$P(D) = \prod_{a_i \in A} \left(\frac{Count(a_i, D) + \mu \cdot P(a_i|R)}{Count(a_\bullet, D) + \mu} \right) \qquad [4]$$

$P(a_i|R)$ est estimée en utilisant le maximum de vraisemblance :

$$P(a_i|R) = \frac{Count(a_i, R)}{Count(a_\bullet, R)} \quad [5]$$

– $P(D)$ représente la probabilité a priori de D.

– $Count(a_i, D)$ représente le nombre d'apparition de l'action a_i dans D.

– $Count(a_i, R)$ représente le nombre d'apparition de l'action spécifique a_i dans la collection R.

– $Count(a_\bullet, X)$ représente le nombre total de signaux sociaux dans X (X est soit le document D ou la collection R).

Ce simple comptage de signaux peut privilégier les "anciennes" ressources par rapport aux récentes, parce que les ressources ayant une longue durée de vie dans le web ont de fortes chances d'obtenir plus de signaux par rapport à celles qui sont récentes. En outre, nous supposons que les ressources qui possèdent des signaux récents sont plus susceptibles d'intéresser les utilisateurs. Nous proposons de prendre en compte les dates associées aux signaux et la ressource. Enfin, nous supposons également que la diversité des signaux sur une ressource est un indice qui peut dénoter un intérêt qui dépasse un réseau social particulier ou une communauté. Nous décrivons dans ce qui suit la manière de prendre en compte ces trois aspects dans l'estimation de $P(D)$.

3.3.1. *Prise en compte de la date de signal*

Nous supposons que les ressources associées aux signaux frais (récents) devraient être favorisées par rapport à ceux qui sont associées à des signaux anciens. Chaque fois qu'un signal apparaît, il est associé à sa date d'occurrence. Nous proposons de compter les occurrences d'un signal en le pondérant (en le *boostant*) avec sa date d'apparition, soit $Count_{t_a}$. La formule correspondante est la suivante :

$$Count_{t_a}(t_{j,a_i}, D) = \sum_{j=1}^{k} f(t_{j,a_i}, D) \quad [6]$$

La pondération de l'occurrence peut se faire de différentes manières. Une façon simple est de prendre par exemple une fonction linéaire :

$$f(t_{j,a_i}, D) = \frac{1}{t_{actuel} - t_{j,a_i}} \quad [7]$$

Ou une pondération exponentielle qui boosterait d'avantage les signaux "récents" vis-à-vis les "vieux" signaux :

$$f(t_{j,a_i}, D) = \exp\left(-\frac{\| t_{actuel} - t_{j,a_i} \|^2}{2\sigma^2}\right) \quad [8]$$

Afin d'éviter la division par zéro qui peut être provoquée par la formule 7, nous utilisons la formule 8 où $f(t_{j,a_i}, D)$ représente la fonction temporelle du signal, estimée en utilisant le noyau Gaussien (Vert *et al.*, 2004). Cette fonction calcule la distance temporelle entre la date actuelle t_{actuel} et la date de l'action t_{j,a_i}. $\sigma \in R_+$ est le paramètre du noyau Gaussien.

Nous notons que plus la distance euclidienne relative au temps $\| t_{actuel} - t_{j,a_i} \|^2$ augmente, plus la valeur du noyau Gaussien diminue. Par conséquent, les signaux sociaux les plus récents sont les plus favorisés.

La probabilité a priori $P(D)$ est estimée en utilisant la formule 4 mais en remplaçant le $Count()$ par $Count_{t_a}()$. Nous notons que si la date du signal est ignorée dans le calcul $f(t_{j,a_i}, D) = 1 \; \forall t_{j,a_i}$.

3.3.2. *Prise en compte de la date de publication de la ressource*

La date de publication d'une ressource joue un rôle important sur la vie sociale de cette ressource dans les réseaux sociaux, c'est-à-dire qu'une vieille ressource a une plus grande chance d'avoir un grand nombre d'interactions par rapport à une ressource publiée récemment. Donc, pour résoudre ce problème, nous proposons de normaliser la distribution des signaux sociaux associés à une ressource par la date de publication de la ressource. On divise le nombre de signaux par l'âge actuel de la ressource. La formule correspondante est la suivante :

$$Count_{t_D}(a_i, D) = \frac{Count(a_i, D)}{Age(D)} \qquad [9]$$

Où :

$$Age(D) = \exp\left(-\frac{\| t_{actuel} - t_D \|^2}{2\sigma^2}\right) \qquad [10]$$

– $Age(D)$ représente la fonction temporelle du document, estimée en utilisant le noyau Gaussien pour les mêmes raisons citées précédemment. Cette fonction calcule la distance temporelle entre la date actuelle t_{actuel} et la date de la ressource t_D.

– Paramètre du noyau Gaussien $\sigma \in R_+$.

La probabilité a priori $P(D)$ est estimée en utilisant la formule 4 mais en remplaçant le $Count()$ par $Count_{t_D}()$ pour le document et $Count_{t_R}()$ pour la collection.

3.3.3. *Prise en compte de la diversité des signaux sociaux dans la ressource*

La diversité et la distribution (répartition) quantitative des signaux sociaux au sein d'une ressource peuvent être des facteurs de pertinence, c.-à-d. une ressource dominée par un seul signal doit être défavorisée par rapport à une ressource ayant une équi-répartition des signaux. On propose d'évaluer cette diversité en utilisant l'indice de diversité de Shannon-Wiener (Pielou, 1966). Cet indice est introduit en écologie pour mesurer la biodiversité. Il est donné par la formule suivante :

$$Diversite_s(D) = -\sum_{i=1}^{m} P(a_i) \cdot log(P(a_i)) \qquad [11]$$

Où : $P(a_i)$ est défini dans la section 3.3

L'indice de Shannon est souvent accompagné par l'indice d'équitabilité de Piélou (Pielou, 1966) :

$$Diversite_s^{Equit}(D) = \frac{Diversite_s(D)}{MAX(Diversite_s(D))} = \frac{Diversite_s(D)}{log(m)} \qquad [12]$$

Avec m représente le nombre de signaux sociaux étudiés. La probabilité a priori $P(D)$ est estimée en utilisant la formule 4 multipliée par le facteur de diversité. La formule correspondante est la suivante :

$$P(D) = \left(\prod_{a_i \in A} P(a_i) \right) \cdot Diversite_s^{Equit}(D) \qquad [13]$$

4. Évaluation expérimentale

Afin de valider notre approche, nous avons effectué une série d'expérimentations sur la collection IMDb (Internet Movie Database). Notre objectif principal dans ces expériences est d'évaluer l'impact des signaux sociaux diversifiés et sensibles au temps sur le système de RI vis-à-vis à la fois d'approches qui ne prennent pas en compte ces facteurs et aussi celles qui ne considèrent pas du tout cette notion de probabilité a priori.

4.1. *Description de la collection de test*

Nous avons utilisé une collection de documents fournies par INEX IMDb. Chaque document décrit un film, et est représenté par un ensemble de méta-données, 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 grammaticaux sont supprimés. Pour chaque document, nous avons collecté des données sociales spécifiques via leur API correspondante sur cinq réseaux sociaux listés dans le tableau 3. Nous les avons mises dans la balise UGC (User Generated Content). Ce champ n'a pas été indexé. Nous avons choisi 30 requêtes parmi l'ensemble des requêtes d'INEX IMDb[1] (voir tableau 2). Pour obtenir

1. https ://inex.mmci.uni-saarland.de/tracks/dc/2011/

les jugements de pertinence, nous utilisons les Qrels fournies par INEX IMDb 2011. Dans notre étude, nous nous sommes concentrés sur l'efficacité des 1000 premiers documents retournés.

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 différentes données sociales	-

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

Requête	Description	Narrative
action biker	search for all action movies with bikers in it.	As i like action movies, specially if bikers are in it, i like to get a list of all these movies.
ancient Rome era	find the movies about the era of ancient Rome.	I am interested in the movies about era of ancient Rome. I am looking for movies talking stories in the era of ancient Rome.
true story drugs +addiction -dealer	find movies about drugs (drug addiction but not drug dealers) that are based on a true story.	I am working with teens and I want to show them a movie about drugs that is based on a true story. A relevant movie is any true story based movie about drug use and addiction. Movies about drug dealers are not relevant. I would like to see as much information as possible about the movie in order to decide whether the movie is appropriate or not.

Tableau 2. *Exemple de requêtes d'évaluation INEX IMDb*

Le tableau 3 montre un exemple de données sociales pour deux documents d'IMDb. L'URL du document est donnée par la syntaxe suivante : *www.imdb.com/title/{id}/*

	Facebook			Google+	Delicious	Twitter	LinkedIn
Id	J'aime	Partage	Commentaire	+1	Bookmark	Tweet	Partage
tt1730728	31	11	2	0	0	2	0
tt1922777	14763	13881	22914	341	12	2859	14

	Facebook		
Id	Dernier Partage	Dernier Commentaire	Date de publication de la Ressource
tt1730728	2013-09-11T20 :55 :47	2012-03-01T11 :07 :32	2010-09-29T05 :08 :09
tt1922777	2014-09-29T02 :49 :01	2014-09-28T00 :41 :01	2011-05-07T19 :00 :57

Tableau 3. *Exemple de deux documents ayant des données sociales*

Malheureusement, les dates des différentes actions ne sont pas disponibles, sauf les dates des dernières actions issues de Facebook (*commentaire* et *partage*). Par conséquent, nous représentons les résultats en utilisant la formule 6 biaisée uniquement par la date du dernier *commentaire* et *partage*.

Réseaux Sociaux	Facebook			Google+	Delicious	Twitter	LinkedIn
Signaux Sociaux	J'aime	Partage	Commentaire	+1	Bookmark	Tweet	Partage(LIn)
Minimum	0	0	0	0	0	0	0
Maximum	76842	43918	62281	1475	986	12223	299880
Total	2478498	2718918	2845169	73392	26143	499232	42787
Moyenne	85.8027	94.1258	98.4964	2.5407	0.9050	17.2830	1.4812

Tableau 4. *Statistiques sur le nombre de signaux sociaux dans les documents retournés par les 30 requêtes*

Réseaux Sociaux	Facebook			Google+	Delicious	Twitter	LinkedIn
Signaux Sociaux	J'aime	Partage	Commentaire	+1	Bookmark	Tweet	Partage(LIn)
AVEC signaux sociaux	16903	18656	13001	5259	3256	12390	3724
SANS signaux sociaux	13097	11344	16999	24741	26744	17610	26276

Tableau 5. *Statistiques sur le nombre de documents (retournés par les 30 requêtes) contenant ou pas des signaux sociaux*

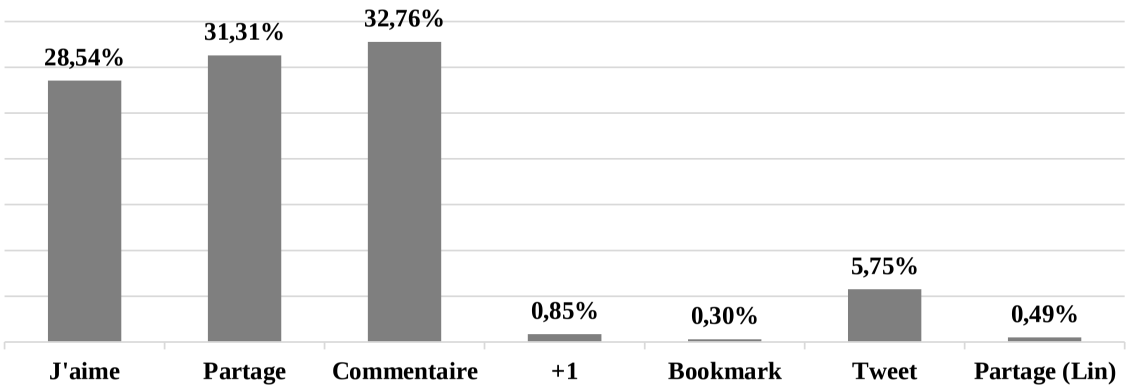

Figure 1. *Pourcentage de distribution de chaque signal dans les documents retournés par les 30 requêtes*

Le tableau 4 présente des statistiques sur le nombre de signaux sociaux dans les documents retournés par les 30 requêtes. Nous remarquons que la densité des signaux de Facebook est très élevée par rapport aux autres signaux. Le tableau 5 présente des statistiques sur le nombre de documents (retournés par les 30 requêtes) contenant ou pas des signaux sociaux. Il y a également la figure 1 qui illustre en pourcentage la distribution de chaque signal dans les documents retournés par les 30 requêtes.

4.2. *Modèles de base*

Nous avons utilisé le moteur de Lucene Solr[2] pour l'indexation et la recherche. Nous avons utilisé le modèle par défaut de Lucene Solr, le modèle de langue Hiemstra (Hiemstra, 1998) et une configuration de notre approche, qui ne prend pas en compte la date de l'action et la date de publication de la ressource, comme des modèles de base.

4.2.1. *Lucene Solr*

C'est un moteur de recherche populaire développé par *Apache Software Foundation* qui est basé sur le modèle vectoriel et la pondération du terme TF-IDF.

4.2.2. *Modèle de langue (ML.Hiemstra)*

Il désigne un modèle d'appariement de RI classique qui utilise le principe de génération de la requête par le document. Le modèle de langue est utilisé dans notre cas pour calculer le score basé sur le contenu textuel.

4.2.3. *Configuration sans prendre en considération le temps*

Afin de pouvoir comprendre l'impact de la diversité et du temps sur les signaux sociaux, nous utilisons également comme modèles de base les différentes configurations qui prennent en compte les signaux sociaux indépendamment du temps et de leur diversité, comme des probabilités a priori de la ressource.

4.3. *Résultats et discussions*

Nous avons mené des expériences avec des modèles basés uniquement sur le contenu textuel des documents (le modèle de Lucene Solr et le modèle de langue Hiemstra sans la probabilité a priori (Hiemstra, 1998)), ainsi que des approches combinant le contenu textuel et les caractéristiques sociales avec prise en compte de leur diversité et de leur aspect temporel. Nous notons que la meilleure valeur de μ appartient à l'intervalle suivant : $\mu \in [90, 100]$.

Pour évaluer la performance de notre approche, nous calculons les métriques de précision (P@k, MAP) et le nDCG (Normalized Discounted Cumulative Gain).

2. http ://lucene.apache.org/solr/

Modèles	P@10	P@20	nDCG	MAP
Base (A) : Sans Probabilité a Priori				
Lucene Solr	0.3411	0.3122	0.3919	0.1782
ML.Hiemstra	0.3700	0.3403	0.4325	0.2402
Base (B) : Sans Prise en Compte de Diversité et du Temps				
J'aime	0.3938	0.3620	0.5130	0.2832
Partage	0.4061	0.3649	0.5262	0.2905
Commentaire	0.3857	0.3551	0.5121	0.2813
TotalFacebook	0.4209	0.4102	0.5681	0.3125
Tweet	0.3879	0.3512	0.4769	0.2735
+1	0.3826	0.3468	0.5017	0.2704
Bookmark	0.3730	0.3414	0.4621	0.2600
Partage (LIn)	0.3739	0.3432	0.4566	0.2515
TousLesCritères	0.4408	0.4262	0.5974	0.3300
a) Avec Prise en Compte de la Date de l'Action T_a				
Partage^{T_a}	0.4148*	0.3681*	0.5472*	0.2970*
Commentaire^{T_a}	0.3861*	0.3601*	0.5207*	0.2844*
b) Avec Prise en Compte de la Date de Publication de la Ressource T_D				
J'aime^{T_D}	0.4091*	0.3620*	0.5308*	0.2907*
Partage^{T_D}	0.4177*	0.3721*	0.5544*	0.2989*
Commentaire^{T_D}	0.3912*	0.3683*	0.5285*	0.2874*
TotalFacebook^{T_D}	0.4302	0.4258	0.5827	0.3200
Tweet^{T_D}	0.3918*	0.3579*	0.4903*	0.2779*
+1^{T_D}	0.3900	0.3511	0.5246	0.2748
Bookmark^{T_D}	0.3732	0.3427	0.4671	0.2618
Partage^{T_D} (LIn)	0.3762	0.3449	0.4606	0.2542
TousLesCritères^{T_D}	0.4484*	0.4305*	0.6200*	0.3366*
c) Avec Prise en Compte de Diversité				
TotalFacebook^{Diversite}	0.4227*	0.4187*	0.5713*	0.3167*
TousLesCritères^{Diversite}	0.4463*	0.4318*	0.6174*	0.3325*
d) Avec Prise en Compte de Diversité et la Date de Publication de la Ressource				
TotalFacebook$^{Diversite}_{T_D}$	0.4417*	0.4289*	0.5966*	0.3273*
TousLesCritères$^{Diversite}_{T_D}$	0.4568*	0.4334*	0.6311*	0.3427*

Tableau 6. *Résultats de P@k, nDCG et MAP*

Le tableau 6 récapitule les résultats de précision@k pour $k \in \{10, 20\}$ et de nDCG, ainsi que la MAP. Nous avons évalué notre approche à travers des configurations différentes, en prenant en compte les signaux sociaux séparément et avec prise en compte de : a) leur date de création ; b) la date de publication de la ressource ; c) leur diversité et d) leur combinaison avec les deux derniers critères. Nous avons déjà montré que la prise en compte de ces signaux sociaux indépendamment de la diversité et du temps améliore significativement la RI par rapport aux modèles basés

uniquement sur la pertinence thématique (Badache et Boughanem, 2014a), (Badache et Boughanem, 2014b). Afin de vérifier si les résultats obtenus sont statistiquement significatifs par rapport aux modèles de base, nous avons effectué le test de Student (Gosset, 1908). Les résultats (*) dans le tableau 6 indiquent que les améliorations sont statistiquement significatives avec un valeur-p (*p-value*) < 0.05. Nous discutons dans ce qui suit les résultats de chacune des configurations que nous avons étudiées.

a) Prise en compte de la date de l'action : le tableau 6.a présente les résultats obtenus en intégrant la date de l'action, dans notre cas, les dates du dernier *commentaire* et *partage* sur Facebook. Les résultats montrent que le nDCG, la précision@k et la MAP sont en général légèrement meilleurs par rapport à ceux obtenus lorsque le temps de l'action est ignoré (Base (B)), mais les résultats restent comparables et significatifs par rapport aux modèles textuels (base (A)). En conséquence, ces résultats confirment partiellement notre hypothèse, que les ressources associées aux signaux frais devraient être favorisées par rapport à ceux associées aux anciens signaux. Cependant, nous n'avons pas vraiment évalué l'impact réel de notre proposition en raison du manque de données appropriées (dates des différentes actions). L'exploitation de la date de la dernière action n'est pas suffisante pour tirer des conclusions efficaces.

b) Prise en compte de la date de publication de la ressource : nous étudions la performance de la RI par l'intégration de la date de publication de la ressource. Les résultats présentés dans le tableau 6.b montrent que le nDCG et les scores de précisions sont meilleurs par rapport à ceux obtenus lorsque la date de publication de la ressource est ignorée (modèles de base (A) et (B)). En effet, une ressource qui génère une activité sociale de 100 actions de *J'aime* pendant une heure de temps, ne suscite pas la même importance et le même intérêt temporel pour les utilisateurs par rapport à une ressource qui a réuni 100 actions de *J'aime* durant une semaine.

c) Prise en compte de la diversité des signaux : le tableau 6.c présente les résultats obtenus par l'intégration de la diversité des signaux au sein du document. Tout d'abord, nous avons effectué une analyse de corrélation en utilisant le coefficient de corrélation de Spearman entre les scores de diversité des signaux dans les documents et leur pertinence. Nous avons trouvé que la diversité représente une corrélation positive avec la pertinence (rho de Spearman = 0.19), ce résultat a suscité notre intérêt et nous a incité à exploiter la diversité dans notre modèle de RI. Le tableau 6.c montre que le nDCG et les précisions sont en général meilleurs que le nDCG et les scores de précision lorsque la diversité est ignorée (modèles de base), mais ces résultats restent faibles par rapport aux résultats où on prend en compte l'âge de la ressource. En effet, si plusieurs utilisateurs de différents réseaux sociaux ont trouvé qu'une ressource est utile, alors il est plus probable que d'autres utilisateurs la trouveront utile aussi.

d) Prise en compte de la diversité et la date de publication de la ressource : le tableau 6.d montre que parmi toutes nos expériences, les meilleurs résultats sont obtenus par la configuration $\textit{TousLesCritères}_{T_D}^{Diversite}$ qui prend en considération la diversité et la date de publication. Par conséquent, le facteur de l'âge de la ressource devient plus efficace quand il est combiné avec la diversité des signaux, ce qui mène à une amélioration significative des résultats comparativement aux modèles de base.

4.4. *Analyse quantitative et qualitative des signaux sociaux*

Afin de mieux comprendre l'effet de ces signaux sociaux sur le processus de sélection des documents pertinents, nous analysons leur distribution dans les différents documents renvoyés pour les 30 requêtes.

	Documents pertinents contenants des signaux			Documents pertinents sans signaux	Documents non-pertinents	
	Nombre de documents	Nombre d'actions	Moyenne	Nombre de documents	Nombre d'actions	Moyenne
J'aime	2210	800458	362.1981	555	1678040	61.6133
Partage	2357	856009	363.1774	408	1862909	68.4012
Commentaire	1988	944023	474.8607	777	1901146	69.8052
Tweet	1735	168448	97.0884	1030	330784	12.1455
+1	790	23665	29.9556	1975	49727	1.8258
Bookmark	429	5654	13.1794	2336	20489	0.7523
Partage (LIn)	601	**40446**	67.2985	2164	**2341**	0.0859
Total	Documents pertinents : 2765			Documents non-pertinents : 27235		

Tableau 7. *Statistiques sur la distribution des signaux dans les documents (pertinents et non-pertinents) retournés par les 30 requêtes*

Le Tableau 7 illustre la distribution des différents signaux dans l'ensemble des documents (pertinents et non-pertinents) renvoyés pas les 30 requêtes. En analysant ce tableau, nous remarquons que la fréquence moyenne des signaux dans les documents pertinents est plus élevée par rapport aux documents non-pertinents (ex. la moyenne des *J'aime* est de 362 actions dans les documents pertinents alors que dans les documents non-pertinents est de 61 actions). Nous remarquons également que les signaux de Facebook capturent la majorité des documents pertinents (voir figure 3) sachant qu'ils sont aussi nombreux dans les documents non-pertinents mais avec une moyenne beaucoup plus petite. Ceci est dû au taux d'engagement sur Facebook et à sa croissance dynamique (Alonso et Kandylas, 2014). Donc, la distinction entre les documents pertinents et les documents non-pertinents est sensible beaucoup plus à la fréquence du signal, c-à-d que les documents pertinents sont caractérisés par un nombre très élevé de signaux Facebook par rapport aux documents non-pertinents (voir figure 2).

Les signaux *Tweet* et *+1* viennent en seconde position avec une fréquence moyenne respectivement de 97 et 29 actions dans les documents pertinents (voir figure 3). Le signal issu de Delicious (*Bookmark*) est le critère le plus faible parmi ces signaux, il n'est présent que dans 429 documents pertinents avec une fréquence moyenne de 13 actions par document seulement. Pour le signal issu de LinkedIn, nous remarquons que 95% de ses actions de *partage* sont concentrées dans 601 documents pertinents avec une fréquence moyenne de 67 actions. Le nombre de documents pertinents capturés par ce signal est très faible par rapport aux signaux de Facebook. Ceci est dû au taux d'engagement sur LinkedIn qui est très faible par rapport à Facebook (Alonso et Kandylas, 2014), mais le signal *partage(LIn)* représente la source la plus fiable en termes de confiance par rapport aux autres signaux sociaux. Par conséquent, la présence de ce signal dans un document représente un indice de pertinence.

Enfin, selon cette étude statistique préliminaire, nous avons constaté que chaque réseau social a sa propre influence sur la qualité de ses signaux. La qualité des signaux,

provenant de Facebook, Twitter, Google+ et Delicious, dépend de leur fréquence, plus les signaux sont fréquents sur le document, plus son importance a priori augmente. Cependant, le signal de LinkedIn ne dépend pas uniquement de sa fréquence parce qu'il a en lui-même une puissance de confiance mature par rapport aux autres signaux. Ceci revient à la maturité des utilisateurs de LinkedIn qui sont mieux réputés par rapport à d'autres utilisateurs des réseaux sociaux.

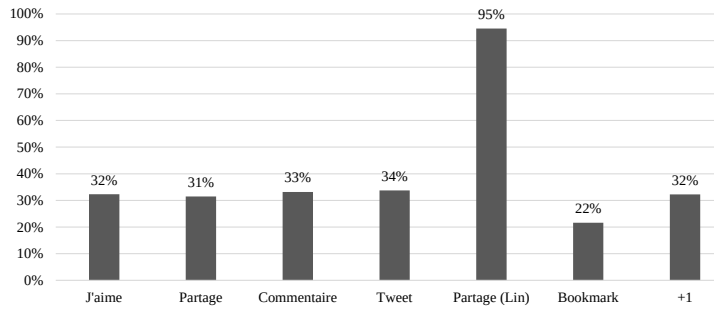

Figure 2. *Pourcentage des signaux dans les documents pertinents*

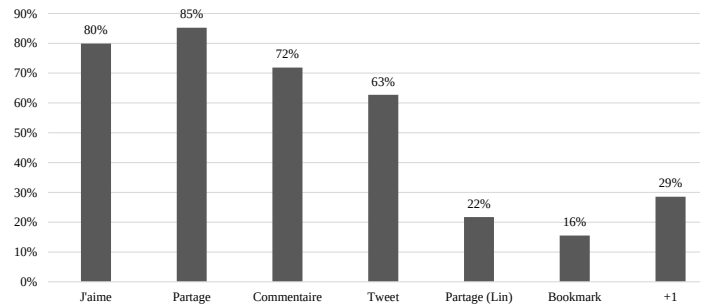

Figure 3. *Pourcentage des documents pertinents contenants des signaux*

5. Conclusion

Dans cet article, nous avons étudié l'impact de la temporalité et la diversité des signaux associés à une ressource, ainsi que la date de publication de cette ressource sur la performance d'un système de RI. Nous avons proposé d'estimer les probabilités a priori du document en tenant compte de ces trois facteurs. Les expérimentations menées sur la collection de données IMDb montrent que la prise en compte des caractéristiques sociales, leur diversité et leurs aspects temporels dans un modèle textuel améliore la qualité des résultats de recherche renvoyés. La contribution principale de ce travail est de montrer que ces facteurs socio-temporels sont fructueux pour les systèmes de RI. Une question importante que nous n'avons pas abordée est l'exploitation du temps associé à chaque action. Malheureusement, actuellement les APIs des réseaux sociaux ne permettent pas l'extraction de ces informations.

Pour les travaux futurs, nous prévoyons de répondre à certaines limites et intégrer d'autres données sociales dans l'approche proposée. D'autres expérimentations sur un autre type de collection sont également nécessaires.

6. Bibliographie

Alonso O., Kandylas V., « A Study on Placement of Social Buttons in Web Pages », *arXiv*, 2014.

Angel A., Koudas N., « Efficient diversity-aware search », *SIGMOD*, ACM, p. 781-792, 2011.

Badache I., Boughanem M., « Harnessing Social Signals to Enhance a Search », *WIC*, vol. 1, IEEE, p. 303-309, 2014a.

Badache I., Boughanem M., « Social Priors to Estimate Relevance of a Resource », *IIiX Conference*, IIiX'14, ACM, NY, USA, p. 106-114, 2014b.

Chelaru S. V., Orellana-Rodriguez C., Altingovde I. S., « Can Social Features Help Learning to Rank Youtube Videos ? », *WISE*, Berlin, p. 552-566, 2012.

Davison B. D., « Topical locality in the Web », *SIGIR*, ACM, p. 272-279, 2000.

Gosset W. S., « The Probable Error of a Mean », *Biometrika*, vol. 6, n° 1, p. 1-25, March, 1908.

Hiemstra D., « A Linguistically Motivated Probabilistic Model of Information Retrieval », *ECDL Conference*, vol. 1513 of *Lecture Notes in Computer Science*, p. 569-584, 1998.

Inagaki Y., Sadagopan N., Dupret G., Dong A., Liao C., Chang Y., Zheng Z., « Session Based Click Features for Recency Ranking. », *AAAI*, vol. 10, p. 1334-1339, 2010.

Karweg B., Hütter C., Böhm K., « Evolving social search based on bookmarks and status messages from social networks », *CIKM*, ACM, p. 1825-1834, 2011.

Kazai G., Milic-Frayling N., « Effects of Social Approval Votes on Search Performance », *Information Technology : New Generations, ITNG '09.*, p. 1554-1559, 2009.

Khodaei A., Alonso O., « Temporally-Aware Signals for Social Search », *SIGIR Workshop on Time-aware Information Access*, 2012.

Khodaei A., Shahabi C., « Social-Textual Search and Ranking », *WWW 2012 CrowdSearch workshop*, 2012.

Macdonald C., Ounis I., « Voting for candidates : adapting data fusion techniques for an expert search task », *CIKM*, ACM, p. 387-396, 2006.

Page L., Brin S., Motwani R., Winograd T., « The PageRank citation ranking : Bringing order to the Web », *WWW*, Brisbane, Australia, p. 161-172, 1998.

Pielou E., « Shannon's formula as a measure of specific diversity : its use and misuse », *American Naturalist*p. 463-465, 1966.

Ponte J. M., Croft W. B., « A Language Modeling Approach to Information Retrieval », *SIGIR'98*, ACM, USA, p. 275-281, 1998.

Qin L., Yu J. X., Chang L., « Diversifying top-k results », *VLDB Endowment*, vol. 5, n° 11, p. 1124-1135, 2012.

Sullivan D., « What Social Signals Do Google & Bing Really Count ? », *Search Engine Land*, 2010.

Vert J.-P., Tsuda K., Schölkopf B., « A primer on kernel methods », *Kernel Methods in Computational Biology*p. 35-70, 2004.

Westerveld T., Kraaij W., Hiemstra D., « Retrieving web pages using content, links, urls and anchors », 2002.

Zhai C., Lafferty J., « A Study of Smoothing Methods for Language Models Applied to Information Retrieval », *ACM Trans. Inf. Syst.*, vol. 22, n° 2, p. 179-214, April, 2004.