Détection des intérêts et de leurs tendances pour des usagers sur des plateformes de *social bookmarking*

Madalina MITRAN, Guillaume CABANAC, Mohand BOUGHANEM
{mitran, cabanac, boughanem}@irit.fr
IRIT/SIG – UMR 5505 CNRS — Université Toulouse 3 – 118 route de Narbonne – F-31062 Toulouse cedex 9

Mots clefs :
social bookmarking, étiquette, recherche d'information, intérêts des usagers, aspect temporel, régression linéaire.
Keywords:
social bookmarking, tag, information retrieval, user interests, temporal aspect, linear regression.
Palabras clave :
social bookmarking, etiqueta, recuperación de información, intereses de los usuarios, aspecto temporal, regresión liniear.

Résumé

Cet article traite des activités des individus qui utilisent une plateforme de *social bookmarking*. Nous proposons d'identifier les intérêts des utilisateurs en intégrant la dimension temporelle et la dimension thématique liée aux *tags*. Plus précisément, l'objectif est de faciliter la recherche des ressources en se basant sur les interactions que les usagers entreprennent sur de telles plateformes. Cette approche repose sur l'identification des individus qui ont des intérêts similaires pour trouver les tendances de leurs intérêts.

Abstract

This paper deals with the activities of individuals who use a social bookmarking platform. Our proposal is to identify the interests of users integrating both time and topical dimensions. We focus on facilitating retrieval of resources, by relying on the interactions that users undertake on such platforms. Our approach is based on the identification of the individuals who have similar interests, and the observation of trends in their interests.

1 Introduction

Depuis l'apparition des systèmes des réseaux sociaux les perceptions des individus pour les ressources deviennent de plus en plus importantes pour le processus de recherche d'information. Nos travaux visent principalement les systèmes de *social bookmarking*.

Dans ce contexte, observer les activités des utilisateurs afin de trouver des similarités entre les individus, de faire des recommandations, ainsi améliorer le processus de recherche d'information est une tâche très difficile. De nos jours, la composante sociale du web est devenue très importante et le besoin de collaboration par l'intermédiaire d'internet a donné naissance à beaucoup de plateformes qui sont des plus en plus utilisés où les utilisateurs peuvent stocker, classer, chercher et partager leurs liens favoris. Ainsi, les internautes ayant les mêmes centres d'intérêts peuvent consulter et sauvegarder les ressources que d'autres ont trouvé. Les services de *social bookmarking* permettent aux utilisateurs d'organiser les ressources en utilisant des étiquettes pour faciliter une recherche ultérieure. Le succès immédiat que de tels systèmes ont eu est dû au fait qu'aucune compétence de la part de l'individu n'est nécessaire pour les utiliser et pour obtenir des bénéfices immédiats.

Dans un premier temps, nous présenterons dans la section 2 une étude des plateformes de *social bookmarking*, le diagramme conceptuel général pour telles plateformes, ainsi qu'une synthèse et les inconvenantes que nous avons identifiés en analysant les plateformes de *social bookmarking*. Ensuite, dans la section 3 nous présenterons notre proposition et les avantages des différents contextes pris en considérations. Ceci nous permettra de discuter, des différentes approches que nous avons considérées en prenant en compte la régression linéaire et la dimension temporelle. Dans la section 4 nous évoquons les principales travaux d'expérimentations de ces approches. Enfin, nous terminons dans la section 5 par les conclusions ainsi que par les perspectives.

2 Étude des plateformes de *social bookmarking*

Les plateformes de *social bookmarking* facilitent la navigation et l'accès aux informations en rendant plus rapide la recherche d'information sur le Web et favorisent la collaboration, la publication et l'archivage des pages Web en facilitant la création d'un espace personnalisé de stockage d'information. Ainsi que, la publication et l'étiquetage des pages avec les auteurs, et le partage des pages en favorisant la collaboration en utilisant les documents sur l'Internet.

Les utilisateurs qui ont un compte sur un plateforme de *social bookmarking* ont la permission d'ajouter des références vers des ressources. Ils disposent aussi d'une page personnelle sur laquelle leurs bookmarks ou leurs publications sont affichés. Sur cette page, tous les bookmarks ou les publications sont affichés dans un ordre chronologique inverse avec une liste de tous les tags que l'individu a utilisés. En sélectionnant un tag, ils peuvent filtrer les bookmarks de sorte que seulement les bookmarks avec cette étiquette sont affichés.

En mettant l'accent sur le potentiel de la communauté, les plateformes de *social bookmarking* offrent des différentes fonctionnalités qui résultent d'analyses sur les bookmarks :

- l'accès aux bookmarks réalisés par d'autres personnes qui ont les mêmes intérêts que soi ;
- la possibilité de consulter les bookmarks des amis et de les ajouter dans sa propre liste (bibliothèque de bookmarks) ;
- taguer avec des mots spécifiques pour faciliter la recherche d'un article qu'il a ajouté ou qui ont été rajoutés par la communauté.

Plus de 100 plateformes *social bookmarking* se trouvent sur Internet. Parmi ces plateformes, nous pouvons mentionner quatre plateformes. Delicious.com a été développé par Joshua Schachter en 2003 et acquis par Yahoo ! en 2005. L'article [8] a détaillé cette plateforme. CiteULike.org a été développé par l'Université de Manchester depuis novembre 2004. Il organise des liens vers des publications académiques et cible tout particulièrement les scientifiques et les chercheurs [6]. BibSonomy.org a été développé par une l'Université Kassel depuis janvier 2006. L'article [3] a mentionné ce plateforme. Connotea.org est un service en ligne de gestion de références

proposé par la maison d'édition Nature depuis décembre 2004. Aussi, il cible tout particulièrement les scientifiques et les chercheurs.

Nous présentons dans la section suivante un diagramme conceptuel des données manipulées par ces plateformes de *social bookmarking*.

2.1 Diagramme conceptuel général pour les plateformes de *social bookmarking*

Nous avons utilisé la notation UML du diagramme de classes afin de modéliser les concepts, les attributs et les associations identifiés dans les plateformes de *social bookmarking*. Dans la figure 1 nous présentons le schéma général qui réunit les caractéristiques des plateformes de *social bookmarking*.

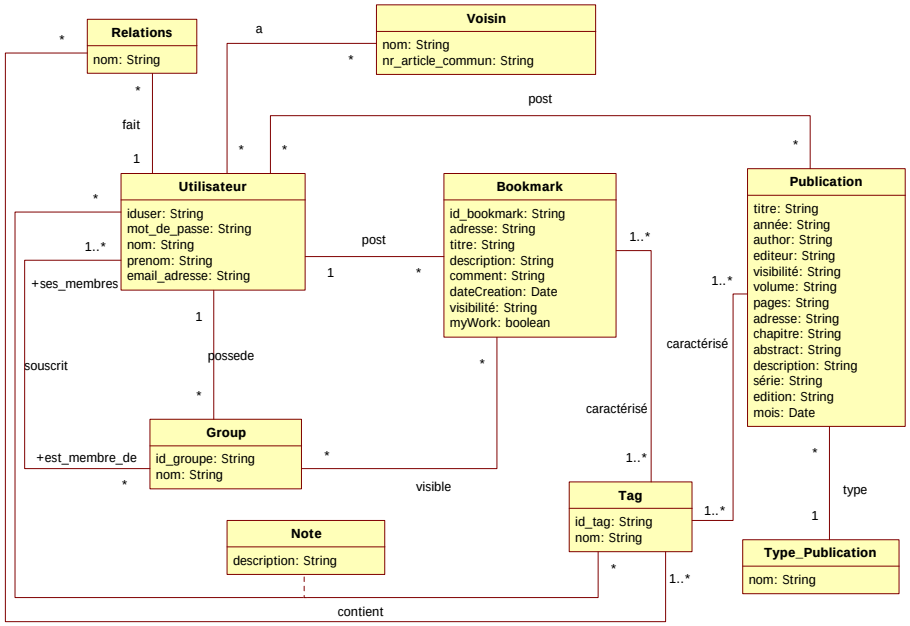


FIGURE 1: Diagramme de classes modélisant les plateformes de *social bookmarking*.

Le schéma contient neuf classes. La classe Utilisateur qui peut être membre de plusieurs groupes, peut avoir zéro ou plusieurs voisins, peut créer plusieurs relations entre les tags. Il peut aussi sauvegarder dans ce compte plusieurs publications et bookmarks. Les classes Bookmark et Publication sont caractérisés par un ou plusieurs tags. Ils sont très similaires, la différence est faite par la classe publication qui a des attributs supplémentaires pour inclure tous les champs BIBTEX. En plus, elle est caractérisée par un seul type (par exemple : article, livre). La classe Relation qui permet à un utilisateur de regrouper des tags. Elle détient l'attribut nom de la relation. La classe Voisin caractérisée par le nom du voisin et le nombre d'articles qu'il a en commun avec l'utilisateur. La classe Tag caractérisée par son nom. La classe Type_Publication caractérisée par son nom. La classe Groupe est caractérisée par les attributs suivants : le nom et la description de groupe, la visibilité (ouvert — les

utilisateurs peuvent trouver le nom de groupe ou la description — et privé — personne ne sera capable de trouver le groupe en cherchant sur le plateforme et les utilisateurs doivent être invités à participer), l'invitation (tous les membres de groupe peuvent envoyer des invitations à rejoindre le groupe ou seulement les administrateurs du groupe peuvent émettre des invitations). De plus, on trouve le droit d'accès (les nouveaux utilisateurs auront le droit d'écrire des articles et de créer des fils de discussion pour lesquels ils auront des droits restreints), l'attribut membre qui permet aux membres d'être anonymes dans leurs activités d'ajouter des ressources, de créer des fils de discussion et poster des commentaires et l'attribut messages, puis d'indiquer si les non-membres peuvent poster ou pas des messages. Enfin, la classe Note est une classe d'association avec un seul attribut qui s'appelle description.

Dans ce schéma nous avons considéré l'intégralité des classes et des attributs qui peuvent illustrer le maximum des fonctionnalités pour les plateformes de *social bookmarking*. En réalité, chaque système particulier adopte une sous-partie de ces fonctionnalités. Nous présenterons dans la section suivante un tableau comparatif des plateformes de *social bookmarking*.

2.2 Synthèse des plateformes de *social bookmarking*

Nous détaillons dans le tableau ci-dessous les caractéristiques des systèmes de *social bookmarking*. Les critères utilisés permettent des comparer les plateformes en considérant le niveau de recommandation (Reco) (des utilisateurs (U) ou des tags (T)), le partage des bookmarks : privé (Pr), public(Pb) ou partiellement visible (PV) ainsi que la notion de *tag cloud*. Il peut être : représentatif pour la communauté (Comm) — les tags les plus fréquents d'un plateforme de *social bookmarking* — ou représentatif pour un seul individu (I) — les tags les plus fréquents d'un individu. Les plateformes figurant dans ce tableau sont présentés dans l'ordre alphabétique de leur nom.

Nom du plateforme	Année	Organisme	Reco		Tag Cloud		Partage		
			U	T	Comm	I	Pb	Pr	PV
Bibsonomy	janvier 2006	Univ. Kassel Germany	−	+	+	+	+	+	+
CiteULike	novembre 2004	Univ. de Manchester	−	+	−	+	+	+	−
Connotea	décembre 2004	Maison d'édition Nature	+	+	+	−	+	+	+
Del.icio.us	septembre 2003	Joshua Schachter	−	+	+	+	+	+	−
Flickr	février 2004	Ludicorp	−	−	+	+	+	+	+

TABLE 1: Comparaison des plateformes de *social bookmarking*.

Sur les plateformes de *social bookmarking*, comprenant des informations sur les individus, les tags et les URLs [11] nous pouvons voir tous les tags employés par un utilisateur (avec leur nombre d'utilisation) depuis son adhésion au plateforme de *social bookmarking* sous la forme d'une liste ordonnée en fonction de la fréquence d'utilisation. En observant leur fonctionnement nous avons identifié certains inconvénients.

1. Concernant l'individu

   a) Impossibilité de voir l'évolution de l'utilisation des tags au cours du temps et implicitement identifier à quel moment une personne a été intéressée par les tags. Par exemple dans une période donnée un individu a été intéressé par le voyage, et particulièrement par un tag, « plage ». Il est impossible de connaître la période d'usage, dans le cas où il l'avait oublié ou s'il est encore intéressé par ce sujet ;

b) Pas de moyen de se concentrer sur l'activité la plus récente d'une personne. Par exemple, pendant 5 ans un individu a été intéressé par la BD, depuis 3 mois il est intéressé par la recherche d'information (RI) (300 tags BD, 30 tags IR). Si quelqu'un veut savoir quel est l'intérêt actuel de l'individu, il peut penser que le BD l'intéresse davantage que la RI ($300 > 30$). Pourtant, il est actuellement plus intéressé par la RI. Donc, le comptage seul ne suffit pas, il faut prendre en compte l'aspect temporel ;

c) Impossibilité de connaître les centres d'intérêts d'un individu sur le long terme et sur le court terme. Par exemple, pendant 4 ans un individu a été intéressé plusieurs fois par les voyages, particulièrement pendant les vacances. Pourtant sur le long terme il est intéressé par le domaine informatique. En regardant, actuellement, l'activité d'une personne sur un plateforme de social bookmarking, nous ne pouvons pas avoir des informations concernant les intérêts sur le long terme et court terme d'une personne. Donc, il faut prendre en compte l'aspect temporel pour pouvoir faire ces observations ;

d) Pas de moyen de trouver des similarités entre les utilisateurs. Par exemple deux utilisateurs ont tagué des ressources avec la même étiquette « recherche » et à peu près avec la même fréquence. Si quelqu'un veux savoir si ces deux utilisateurs ont des intérêts communs, il peut penser qu'ils ont les mêmes intérêts. Pourtant il est possible que ces deux utilisateurs ont été intéressés pour ce tag à différentes périodes du temps et qu'à présent ils ont des intérêts opposés. Donc, les tags utilisés en commun et la fréquence de ces tags ne sont pas suffisant pour trouver des similarités entre les utilisateurs ;

e) Pas de moyen d'observer les cooccurrences entre les tags. Par exemple, si un individu décrit une ressource en choisissant plusieurs tags, nous n'avons pas la possibilité d'observer la tendance pour l'ensemble de ces tags au cours du temps. Peut-être qu'il les a utilisés simultanément seulement un ou deux fois par hasard ou fréquemment. Donc, nous ne pouvons rien dire sur l'intérêt d'un individu porte à ces tags au cours du temps.

f) Pas de moyen d'observer les similarités entre tags pour former des domaines d'intérêt.

2. Concernant la communauté :

a) Impossibilité de connaître les centres d'intérêts de la communauté sur le long terme et court terme ;

b) Pas de moyen de trouver des similarités entre les groupes d'utilisateurs ;

Pour l'utilisateur d'un plateforme de *social bookmarking* tous ces inconvénients rendent difficile le processus de recherche d'information. Le usagers ne bénéficient pas pleinement des ressources que les autres ont trouvé et qui peuvent être intéressantes pour des utilisateurs ayant des intérêts communs.

3 Restituer l'activité des usagers en prenant en compte la dimension temporelle

Cette section est divisée en deux parties. Nous présentons, dans la première partie les avantages des différents contextes pris en considérations. Elle vise à répondre à la problématique que nous avons présentée dans la section 2.2. Nous nous intéressons en particulier aux activités des utilisateurs sur un plateforme de *social bookmarking*. Dans la deuxième section, nous présentons les différents approches que nous avons considérées en prenant en compte la régression linéaire et la dimension temporelle.

3.1 Proposition

Selon Dubinko et al. [5], si nous pouvions observer le comportement des utilisateurs au fil du temps, nous pourrions explorer l'évolution des leurs centres d'intérêts qui constituent un avantage pas seulement pour l'individu mais aussi pour la communauté.

Nos propositions visent à améliorer le processus de recherche d'information, à trouver les différents intérêts des individus et à trouver des similarités entre eux, en s'appuyant sur les activités des individus qui utilisent les plateformes de *social bookmarking*, tout en prenant en compte l'aspect temporel. Nous apportons des éléments de réflexion concernant les points faibles des plateformes de *social bookmarking*. Ainsi, nous identifions les avantages pour différents contextes en fonction du nombre de tags et du nombre d'utilisateurs.

Nous identifions des avantages pour différents contextes en fonction du nombre de tags et du nombre des utilisateurs en prenant en compte l'aspect temporel. Ils sont détaillés ci-dessous.

(a) Le contexte individuel relatif à un tag.

Nous proposons une visualisation immédiate de la tendance qu'un tag peut suivre en fonction du temps. En observant ces tendances sur la page d'un individu, nous identifions plusieurs avantages. Premièrement, ceci facilite le travail de recherche des autres utilisateurs, de sorte que si un individu est intéressé par un tag à un moment donné, il peut voir d'un clin d'œil les activités des autres utilisateurs concernant le même tag, afin qu'il puisse bénéficier des ressources que d'autres ont trouvé. De plus, nous pouvons observer facilement l'actualité des ressources et implicitement, trouver des personnes qui ont les mêmes centres d'intérêts ou des personnes qui se rassemblent.

Nous proposons aussi de diviser toute la période d'usage d'un plateforme de *social bookmarking* d'un individu en d'intervalles de temps pour trouver les intérêts constants à long terme et les intérêts constants à court terme. Les avantages de cette proposition concernent les individus qui peuvent trouver des personnes qui ont des intérêts similaires avec eux et pourquoi pas mettre en œuvre une approche de recommandation. En connaissant les intérêts à long terme et à court terme des individus, nous pouvons leur recommander des utilisateurs et des ressources, et ainsi les aider dans leur processus de recherche.

(b) Le contexte organisationnel relatif à un tag.

Ce contexte est similaire au premier contexte que nous avons proposé. La différence entre ces deux est faite par le nombre d'utilisateurs pris en compte. Dans cette deuxième approche nous trouverons des groupes d'individus qui ont les mêmes activités en observant les similarités entre les utilisateurs.

(c) Le contexte personnel relatif à un tag.

Un utilisateur d'un plateforme de *social bookmarking* peut trouver intéressant le fait de pouvoir visualiser son activité concernant un tag en fonction du temps. Ceci lui permet de trouver à quel moment il a été intéressé par un tag. Aussi, nous pouvons trouver la fraîcheur des intérêts de nos amis (connexions).

(d) Les contextes individuel, organisationnel et personnel relatifs à plusieurs tags.

Comme les avantages de ces trois contextes sont très proches les uns des autres, nous préférerons les aborder dans la même section. Nous proposons pour ces contextes deux approches : l'approche basée sur la thématique (le domaine) et l'approche basée sur les trois tags les plus représentatifs (3TPR). Nous détaillons ces deux approches ci-dessous.

(i) L'approche de la thématique

En étudiant les tags cooccurrents (les tags qui sont utilisés ensemble pour les utilisateurs pour étiqueter les ressources) nous formons des domaines d'intérêts (des thématiques) pour les individus en fonction de ces étiquettes. En observant la tendance pour ces domaines dans le temps pour chaque utilisateur nous

trouverons facilement son intérêt pour une certaine thématique. En plus, les utilisateurs qui sont intéressés par les mêmes domaines peuvent bénéficier, sans perdre beaucoup de temps, des ressources que d'autres ont trouvé.

(ii) L'approche de 3TPR prend en compte les trois tags les plus fréquents qui caractérisent le mieux l'utilisateur. Ainsi, en un seul clin d'œil nous pouvons connaître le profil d'un utilisateur, ses centres d'intérêt, les domaines, les sujets qui l'intéressent. En plus, nous gagnons du temps si nous pouvons dire en quelque secondes si les activités d'un individu peuvent être avantageuses ou pas pour nous.

Les limites de ces contextes sont basées sur les inconvénients des tags. Imaginons deux individus, l'un utilisant le tag « recherche » et l'autre le tag « research», nous ne pouvons rien dire sur le deux individus parce que ces deux tags sont considérés différents même s'ils ont la même sémantique.

Dans notre étude, nous avons fait l'hypothèse que les individus d'un plateforme de *social bookmarking* utilisent les mêmes tags en éliminant les inconvénients que les tags peuvent avoir.

Après avoir exposé les avantages que l'aspect temporel peut apporter dans le cadre des plateformes de *social bookmarking* et implicitement dans la recherche d'information, on a besoin d'une représentation visuelle des activités des individus. La représentation d'un tag seulement par des points (voir figure 2) ne suffit pas. Observer la tendance est plus informatif. Cela permet d'examiner plus clairement et plus rapidement les activités des utilisateurs. À cet effet, nous avons utilisé la régression linéaire.

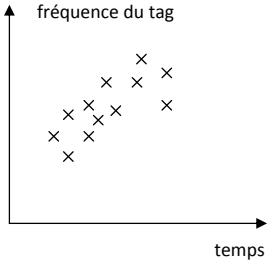


FIGURE 2: Représentation par des points

Dans la suite nous présenterons les approches que nous avons considérées.

3.2 Approches

3.2.1 Identification des trois tags les plus représentatifs

Nous considérons la collection de bookmarks d'un utilisateur qui est sauvegardée sur un plateforme de *social bookmaking*. Chaque bookmark est composé d'un ou plusieurs tags. Pour identifier les 3TPR nous proposons deux approches.

(a) Première approche

Nous considérons la fréquence des tags dans la collection pour identifier l'intérêt d'un utilisateur pour une certaine thématique. Plus un tag est utilisé, plus l'intérêt de l'usager est important. La méthode la plus simple pour déterminer les 3TPR est de compter le nombre d'occurrence de chaque étiquette dans la collection. Pour notre travail de recherche, cette approche se confronté à une limite.

Un individu a pu utiliser souvent un tag donné, sans que ce dernier l'intéresse encore actuellement. Cela signifie que l'intérêt actuel de l'usager pour cette étiquette est en déclin. De cette façon la fréquence d'un tag dans la collection d'étiquettes d'un utilisateur n'est pas suffisante pour constater l'intérêt ou l'indifférence pour un certain sujet. La deuxième approche que nous présentons dans le paragraphe suivant apporte une solution à cette limite introduite par la première approche.

(b) Deuxième approche

Le choix des 3TPR est plus spécifique si nous considérons la tendance d'un tag en fonction de sa fréquence d'apparition dans la collection d'étiquettes d'un utilisateur. Ainsi, pour représenter les véritables intérêts d'un individu, nous prenons en compte deux paramètres : la fréquence et la tendance d'utilisation des tags en fonction du temps.

Nous proposons de calculer, pour chaque tag, un score qui tient compte de ces deux paramètres : $Score(t) = \alpha \cdot freq(t) + (1 - \alpha) \cdot penteTendance$ avec $\alpha \in [0, 1]$.

La fréquence du tag représente le nombre d'occurrences du tag dans l'ensemble des étiquettes d'un individu qui utilise un système de *social bookmarking*. Pour éviter les biais liés à la longueur de la collection (le nombre d'occurrence serait potentiellement plus élevé pour un individu qui utilise beaucoup un plateforme de *social bookmarking* par rapport à un individu qui l'utilise rarement) nous allons normaliser la somme de chaque tag.

Pour normaliser nous assimilons l'utilisateur à un document et un tag à un mot et nous utilisons la mesure tf [14] exprimée par le rapport entre le nombre d'occurrences d'un tag dans la collection et le nombre d'occurrences de tous les tags dans la collection d'un utilisateur. Soit la collection des étiquettes d'un utilisateur c_j et un tag t_i, la fréquence du tag dans la collection des tags (tf) est calculée par la formule suivante : $tf(u, t) = \frac{n_{i,j}}{\sum_k n_{k,j}}$ où $n_{i,j}$ est le nombre d'occurrences du tag t_i dans c_j et le dénominateur est le nombre d'occurrences de tous les tags dans la collection c_j.

Nous observons la tendance d'un tag d'un utilisateur en prenant en compte l'aspect temporel. De cette façon, pour chaque tag, nous déterminons la pente de la régression linéaire. La valeur que nous obtenons pour la pente est intégrée dans la formule de score (fonction $Score(t)$) présentée précédemment. Pour calculer la pente nous utilisons la suivante formule : $penteTendance = \frac{a}{b}$.

Chaque tag qui se trouve dans la collection d'un utilisateur est alors caractérisé par un score. À l'aide de celui-ci, nous ordonnons les tags par score décroissant. À partir de la liste obtenue, nous ne conservons que les trois premiers tags pour représenter l'intérêt de l'utilisateur en fonction du temps.

Dans cette approche nous avons calculé un score pour toute la période dans laquelle l'individu a utilisé un plateforme de *social bookmarking*. En analysant les activités des utilisateurs sur les plateformes du *social bookmarking* nous pouvons faire les remarques suivantes : parmi les 3TPR il y a des tags que l'individu a utilisé dans le passé et pour lesquels il n'est actuellement plus intéressé, il y a des étiquettes qui représentent les intérêts actuels mais qui n'ont pas été utilisées antérieurement et il y a des tags qui ont été utilisés constamment.

En prenant en compte tout la période d'usage, nous ne pouvons pas observer les tags décrits ci-dessus, cela montre une limite de notre approche. Afin de répondre à cette limite, le paragraphe suivante expose une troisième approche.

(c) Troisième approche

Pour que le choix des 3TPR soit plus spécifique et plus représentatif pour l'activité d'un usager, nous proposons de diviser la période d'interaction d'un individu avec un plateforme de *social bookmarking* en plusieurs intervalles de temps.

Nous prenons l'exemple d'un utilisateur (Guillaume Cabanac [1]) qui a utilisé un plateforme de *social bookmarking* depuis 2007 jusqu'à aujourd'hui. Nous divisons la période d'usage en 4 intervalles (pour chaque année).

Nous calculons pour chaque intervalle de temps le score que nous avons défini dans la section 3.2.1 et nous présenterons les résultats dans le tableau 2.

3TPR
information retrieval
dev_JAVA
ea_16_05_2007

(a) 2007

3TPR
information retrieval
dev_JAVA
latex

(b) 2008

3TPR
latex
annotation
information retrieval

(c) 2009

3TPR
annotation
information retrieval
to_Madalina

(d) 2010

TABLE 2: Les tags pour l'activité d'un usager divisés en quatre intervalles.

Diviser la période d'interaction d'un utilisateur avec un plateforme de *social bookmarking* en plusieurs intervalles apporte des avantages. Nous pouvons ainsi mieux observer les intérêts actuels, les intérêts à long terme, les intérêts en déclin et les thématiques pour lesquelles il n'est plus intéressé. En regardant l'exemple que nous avons pris, nous pouvons tirer les observations suivantes :

- le tag ea_16_05_2007 est une étiquette que l'individu a utilisé beaucoup en 2007 et à présent il n'est plus intéressé par ce sujet.
- le domaine information retrieval représente un intérêt à long terme (il est utilisé approximativement avec la même fréquence dans chaque intervalle de temps).
- le tag dev_Java représente un intérêt qui est en déclin. Il a été beaucoup utilisé dans deux intervalles de temps (en particulier en 2007 et 2008). En 2009 et 2010 il est utilisé mais pas avec la même fréquence, ce qui nous fait penser que l'intérêt d'utilisateur pour cette étiquette a baisé.
- nous observons aussi les nouveaux (actuels) intérêts d'un utilisateur, par exemple le tag to_Madalina dans l'intervalle 2010 est une étiquette pour laquelle l'individu s'est beaucoup intéressé récemment.

Cette partie a permis d'identifier les 3TPR en utilisant la fréquence et la régression linéaire pour différents intervalles des temps. Dans la suite nous proposons d'adopter la régression linéaire pour trouver des similarités entre les individus qui interagissent avec un système de *social bookmarking*.

3.2.2 Identification des similarités entre les utilisateurs

Nous proposons deux méthodes pour trouver les similarités entre les individus qui utilisent un système de *social bookmarking*.

1. `http://www.connotea.org/user/Tafanor` sur *Connotea*

a) Première méthode

Pour trouver des similarités entre les utilisateurs, nous observons les tendances de plusieurs individus pour le même tag et pour la même période de temps (figure 3). Nous superposons les graphiques avec ces tendances et nous calculons à l'aide d'une métrique les similarités entre les utilisateurs. Des tendances confondues indiquent que les utilisateurs ont les mêmes centres d'intérêts. Dans le cas contraire, les centres d'intérêt des individus s'éloignent avec la croissance de l'angle entre les tendances. Pour calculer les similarités entre les utilisateurs nous utilisons la mesure du cosinus, tout comme cela a été proposé dans [10].

Nous calculons la similarité entre l'utilisateur 1 et l'utilisateur 2 après la formule suivante : $sim(u_1, u_2) = cos(\vec{u_1}, \vec{u_2}) = \alpha$.

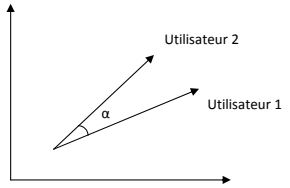


FIGURE 3: Représentation des tendances des deux utilisateurs

Après avoir déterminé les individus qui ont des intérêts similaires, nous faisons des groupes avec ceux-ci. Dans le paragraphe suivant nous décrivons une deuxième méthode pour trouver les similitudes entre les utilisateurs

b) Deuxième méthode

La deuxième méthode que nous utilisons pour trouver les individus qui ont les mêmes centres d'intérêt et former des groupes avec ceux-ci s'appuie sur l'approche des 3TPR.Nous proposons de classifier premièrement les tags dans des domaines. Pour cela nous allons analyser la cooccurrence des tags. Plus la fréquence des tags qui sont utilisés ensemble augmente, plus ces tags tendent à former un domaine.

Cette deuxième méthode consiste en deux étapes. En premier lieu, nous allons construire la matrice des fréquences de cooccurrences des tags. Ensuite, après la normalisation et le tri par blocs diagonaux de cette matrice [1], elle va contenir sur la diagonale principale des blocs correspondants à des classes de tags qui ont été utilisés ensemble par les usagers du plateforme de *social bookmarking* analysé. Ainsi les domaines trouvés, nous décrivons les individus comme : $\vec{u_1} = \{t_1^1, \cdots, t_1^n\}$, $\vec{u_2} = \{t_2^1, \cdots, t_2^m\}$, $\ldots \vec{u_r} = \{t_r^1, \cdots, t_r^q\}$. Chaque individu qui utilise un plateforme de *social bookmarking* est caractérisé par l'ensemble des ses tags.

Pour trouver les similarités entre les utilisateurs, nous prenons pour chacun son 3TPR et nous identifions pour chaque tag son domaine d'appartenance. En mettant les individus dans des groupes en tenant compte des domaines auxquels ils sont intéressés nous identifions les personnes qui se ressemblent (ont des intérêts similaires).

Pour une meilleure compréhension nous présenterons ci-dessous un exemple. Pour notre exemple nous considérons l'ensemble suivant des tags : Tags={inf_retrieval, evaluation, web, visualisation, TREC, indexation, programming, literature, computer, language, application, information, developement, library, cultural, educational, book}. Avec ces tags nous constituons la matrice dont les lignes et les colonnes comportent des étiquettes. Les cellules indiquent la fréquence de cooccurrence des tags. Après la normalisation et le tri de cette matrice nous identifions trois domaines. Ils sont décrits par les ensembles suivants d'étiquettes : D_1={inf_retrieval, evaluation, web, visualisation, TREC, indexation, information},D_2={programming, language, application, information, computer, developement}, et D_3={literature, language, library, cultural, educational, book}. Nous prenons les 3TPR de cinq utilisateurs et nous identifions pour chacun son domaine d'appartenance (tableau 3).

3TPR
inf_retrieval
evaluation
programming

(a) u_1

3TPR
web
literature
book

(b) u_2

3TPR
inf_retrieval
TREC
indexation

(c) u_3

3TPR
educational
developement
book

(d) u_4

3TPR
literature
library
cultural

(e) u_5

TABLE 3: Les 3TPR des utilisateurs.

Nous identifions les groupes des individus en regardant les domaines pour chaque tag. Nous trouvons les classes suivantes d'utilisateurs : $C_1 = \{u_1, u_2, u_3\}$, $C_2 = \{u_1, u_4\}$ et $C_3 = \{u_2, u_4, u_5\}$.

À l'aide de ces deux méthodes nous pouvons faire des recommandations pour les utilisateurs qui se trouvent dans le même groupe. En plus, les individus qui font partie de la même classe peuvent trouvé facilement des ressources que d'autres ont trouvées en diminuant le temps nécessaire de recherche.

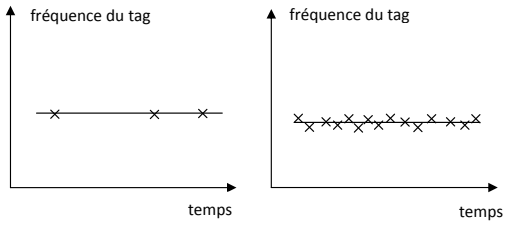
Cette partie a permis d'identifier les similarités entre les utilisateurs en utilisant deux techniques : la mesure du cosinus et la méthode de classification par analyse de connexité. Pour notre travail de recherche ces techniques s'appuient sur la régression linéaire et sur l'aspect temporel. Dans la suite, nous proposons identifier les intérêts constants à long terme et implicitement à court terme pour les individus d'un plateforme de *social bookmarking*.

3.2.3 Identification des intérêts constants sur le long terme et sur le court terme

En identifiant les intérêts récurrents à l'aide de la régression linéaire, nous voulons identifier si ceux-ci représentent des intérêts constants à long terme ou intérêts constants à court terme pour les utilisateurs. Nous identifions deux approches :

(a) Première approche

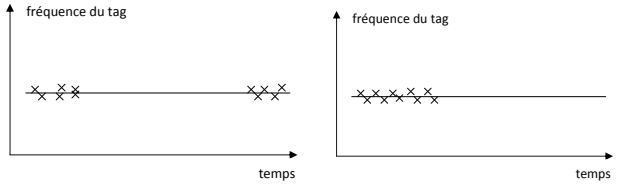
Les intérêts récurrents que nous avons identifiés pour un individu au fil de temps ont des fréquences différentes (par exemple, pour deux tags qui représentent des intérêts constants, un tag a été utilisé 3 fois, figure 4 (a), par rapport au deuxième qui a été utilisé 15 fois dans la même période du temps, figure 4 (b)).



(a) (b)

FIGURE 4: Intérêt récurrent

Nous proposons donc de prendre en considération la fréquence d'apparition pour les tags qui montre un intérêt récurrent en fonction du temps. Cette approche se confronte à une limite dans le cadre de notre travail. En tenant compte de la fréquence d'apparition des étiquettes, nous ne pouvons pas dire si elles représentent des intérêts constants à long terme pour les individus. Nous pouvons trouver des périodes dans lesquelles l'utilisateur n'a pas montré un intérêt pour l'étiquette considérée (figure 5 (a)) ou des étiquettes récurrenes pour lesquels l'individu était beaucoup intéressé seulement dans un période de temps (figure 5 (b)). Pour ces tags nous pouvons dire qu'ils représentent des intérêts constants à court terme plutôt qu'à long terme.



(a) (b)

FIGURE 5: Intérêt récurrent en fonction du temps avec des périodes manquantes (a), Intérêt récurrent à court terme (b)

Dans le paragraphe suivant nous présenterons une deuxième approche qui apporte une solution à cette limite.

(b) Deuxième approche

Notre deuxième approche est basée sur le travaux de Dubinko et al. [5]. Notre proposition pour déterminer les intérêts constants à long terme doit avoir les propriétés suivantes :

1) une étiquette doit être considérée récurrente pour toute la période d'usage d'un plateforme de *social bookmarking* s'il a été constant pendant tous les intervalles des temps.
2) une étiquette qui est récurrente seulement pendant un intervalle de temps particulier ne devrait pas représenter nécessairement un intérêt constant à long terme pour un individu.

Nous divisons toute la période d'utilisation d'un plateforme de *social bookmarking* d'un individu en des intervalles de temps. L'intervalle de temps peut représenter : des mois, des trimestres, des années. Soit $\mathcal{I} = [a, b]$ un intervalle de temps. À l'aide de la régression linéaire nous exprimons pour chaque étiquette dans intervalle $\mathcal{I}$, le fait qu'elle représente un intérêt récurrent ($Récurrent(t, \mathcal{I}) = const$).

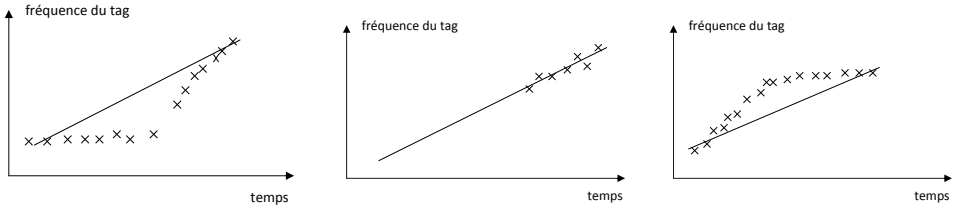
Notre trouverons les intérêts constants à long terme si pour chaque intervalle considéré nous observons pour la même étiquette un intérêt récurrent. Nous exprimons cela par la formule suivante : $Récurrent(t, \mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_n) = \sum_{i=1}^{n} Récurrent(t, \mathcal{I}_i) = const$

Cette partie a permis d'identifier les intérêts constants sur le long terme et implicitement les intérêts constants sur le court terme. Dans la suite nous identifions les étiquettes qui représentent les nouveaux intérêts des usagers et les étiquettes qui, dans le passé, n'ont pas été beaucoup utilisées et qui deviennent maintenant elles pour les individus qui utilisent les plateformes de *social bookmarking*.

3.2.4 Identification des TIM, des NTI et des TIA

Premièrement nous spécifions les concepts des « tags important maintenant » (TIM), des « nouveaux tags intéressantes » (NTI) et des« tags importants avant » (TIA). Ensuite, nous présenterons le coefficient de détermination qui va nous permettre de les identifiés.

Les TIM sont des tags qu'un individu a utilisés dans le passé d'une façon constante et pour lesquels il montre actuellement un intérêt croissant. Les NTI sont les tags que les individus n'ont pas utilisés dans le passé et pour lesquels ils sont beaucoup intéressés actuellement. Les TIA sont les étiquettes que les individus ont beaucoup utilisés dans le passé et pour lesquelles ils montrent un intérêt récurant actuellement. Nous présentons ces phénomènes dans la figure 6.



(a) (b) (c)

FIGURE 6: Graphique pour un TIM (a), Graphique pour un NTI (b) et Graphique pour un TIA (c)

La régression linéaire seule ne suffit pas à déterminer les TIM, NTI et TIA parce que pour les trois nous identifions un intérêt croissant. Dans la section suivante nous présenterons le coefficient de détermination qui apporte une solution à ce problème.

Coefficient de détermination

Dans la régression, le coefficient de détermination $r^2 = cor(x, y)^2$ [4] est une mesure statistique de la façon dont la droite de régression se rapproche des points. Il est égal au carré de coefficient du corrélation entre la variable réponse x et la variable prédictive y. Le coefficient de détermination r^2 satisfait la relation : $0 \leqslant r^2 \leqslant 1$.

Une valeur de 1 indique que la ligne de régression s'inscrit parfaitement dans les données. Dans notre cas, une valeur proche de 1 indique le fait que les étiquettes analysées sont des NTI et pour une valeur proche de 0 nous trouvons les TIM.

Nous avons exposé dans cette section nos approches où nous avons proposé d'utiliser la régression linéaire et de prendre en compte la dimension temporelle pour trouver les intérêts des individus au fil de temps, pour trouver les similarités entre les usagers dans le but d'aider les individu dans leur processus de recherche d'information. Dans la section suivante nous présenterons les démarches que nous envisageons à faire pour évaluer notre travail.

4 Évaluation

Notre travail d'évaluation est basé sur des données expérimentales extraites d'un plateforme de *social bookmarking* Connotea, actuellement utilisé par plus d'un million d'utilisateurs. Il permet de gérer et de partager leurs collections de bookmarks sur le web. L'étude est basée sur le plateforme Connotea pour plusieurs raisons. Elle est un plateforme professionnel utilisé par des personnes du monde scientifique, elle est un plateforme qui a commence à être très populaire et de plus en plus utilisé et parce que il est souhaitable de travailler sur des données fiables, des données de qualité

Nous extrairons du plateforme Connotea les utilisateurs, les tags, les bookmarcks et les informations sur les ressources et ils sont stockés dans un base de données appelée cabanac_connotea. Pour évaluer nos approches nous utilisons une interface graphique et nous faisons des statistiques sur des données que nous obtiendrons à l'aide des utilisateurs.

4.1 Évaluer l'approche des intérêts constants sur le long terme et sur le court terme avec le coefficient de corrélation de Kendall

Nous utilisons dans notre étude le coefficient de corrélation τ de Kendall [2] pour évaluer le degré de similarité entre deux séries ordonnés différemment pour un même ensemble de tags. Le coefficient de Kendall dépend du nombre d'inversion des paires des tags qui seraient nécessaires pour transformer un ordre de classement dans l'autre.

Nous demandons aux utilisateurs du plateforme de *social bookmarking* Connotea d'ordonner de façon ascendante une liste d'étiquettes en fonction de leurs intérêts récurrents les plus représentatifs pour eux à long terme. La liste est formée de tags pour lesquels les individus ont montré des intérêts constants.

Nous avons notre propre liste avec un classement qui tient compte de l'approche présentée dans la section 3.2.3. En fonction de la valeur que nous avons obtenue pour le coefficient τ de Kendall nous pouvons dire si l'approche que nous avons considérée identifie bien les intérêts constant à long terme pour les utilisateurs.

Pour une valeur grande de τ, nous pouvons dire que les deux listes coïncident et que notre approche est performante. Dans le cas contraire nous affirmons que les deux listes sont totalement différentes.

4.2 Évaluer l'approche de 3TPR

Après que nous ayons identifié les trois tags conformément à la section 3.2.1, nous évaluons notre approche en demandant aux utilisateurs de donner des notes aux graphiques en fonctions de leurs intérêts actuels (s'ils représentent ou pas leurs intérêts actuels).

Notre interface d'évaluation contient des graphiques qui représentent les tendances pour différentes étiquettes. Ces graphiques montrent : l'intérêt d'un individu à un moment donné, avec des étiquettes qui ne sont pas utilisées par l'individu mais qui illustrent un intérêt croissant dans le temps, leurs intérêts constants au cours du temps, des intérêts en déclin et avec les 3TPR issus de notre étude.

En fonction des réponses que nous obtenons des utilisateurs, nous pouvons conclure si les 3TPR représentent bien les intérêts actuels des individus. Notons que nous avons intentionnellement bruité les listes présentées aux participants (tags qui n'appartiennent pas aux participants) pour vérifier qu'ils reconnaissent leurs propres tags.

5 Conclusion et perspectives

Le travail présenté dans ce article s'inscrit dans le cadre de l'exploitation des systèmes de *social bookmarking*. Nous avons déterminé l'impact que l'introduction de la dimension temporelle sur un tel plateforme peut avoir dans les activités des individus et implicitement pour la communauté. En plus, nous avons établi l'effet que la division de toute la période d'utilisation d'un plateforme de *social bookmarking* en intervalle de temps permet aux utilisateurs d'avoir une meilleure vision de leur activité. Plus précisément, nous avons étudié l'activité d'étiquetage des ressources par rapport au temps, en utilisant la régression linéaire avec différentes techniques (coefficient de corrélation, coefficient de détermination, méthode de moindres carrées, mesure de cosinus et méthode de classification par analyse de connexité). En observant les interactions que les personnes ont avec un plateforme de *social bookmarking*, nous trouvons les intérêts des utilisateurs (intérêt en déclin, croissant, récurrent ainsi que les intérêts à long terme et à court terme) au cours du temps ainsi que les similarités entre les utilisateurs.

La façon avec laquelle les usagers interagissent avec le plateforme de *social bookmarking* est très différente en fonction du temps. À cause de cela, la régression linéaire ne représente pas toujours de façon optimale les intérêts des utilisateurs. Une alternative qui permet de mieux modéliser de tels intérêts existe. En effet, il y a différentes techniques comme la régression polynomiale, que nous pouvons intégrer dans notre approche.

Comme perspective à notre travail, nous proposons d'utiliser notre approche pour recommander des usagers et des ressources à des personnes qui partagent les mêmes centres d'intérêts au fil de temps. De cette façon nous facilitons la tâche de recherche en réduisant le temps pour trouver des documents pertinents. Plusieurs systèmes de recommandation ont été présentés dans : [13, 15, 16, 9, 12]. Il conviendra alors de se comparer à ces travaux.

Nous envisageons aussi d'étendre notre proposition pour supporter d'autres tâches de la recherche d'information dans les systèmes de *social bookmarking*. Selon la théorie du sociologue Gladwell [7] il existe trois trois types de personnes selon leur rapport à l'information :

- les *connectors* sont caractérisés par leur grand nombre d'accointances, ils arrivent à établir des liens entre différentes communautés ce qui leur permet d'y disséminer l'information.
- les *mavens* accumulent les savoirs, disposent et sont à l'origine de nombreuses informations qu'ils partagent volontiers autour d'eux, dans un cercle réduit d'accointances ;
- les *salesmen* promeuvent les nouvelles idées qu'ils glanent, savent les valoriser et les diffuser autour d'eux.

Ce travail peut être adapté pour permettre d'identifier les différentes caractéristiques que les usagers peuvent avoir. Avec ce type d'information, nous pourrions adapter le processus de recherche d'information en détectant les nouvelles informations et tendances (*mavens*), en les recommandant aux personnes pour les valoriser (*salesmen*), pour accroître leur visibilité et leur dissémination dans les différents cercles d'accointances (*connectors*).

References

[1] Tétralogie : Logiciel de veille scientifique et technique. `http://atlas.irit.fr/index.html`.

[2] H. Abdi. The kendall rank correlation coefficient.

[3] D. Benz, F. Eisterlehner, A. Hotho, R. Jäschke, B. Krause, and G. Stumme. *Managing publications and bookmarks with BibSonomy*. ACM, New York, NY, USA, June 2009.

[4] S. Chatterjee and B. Price. *Regression analysis by example*. A Wiley-Interscience publication. Wiley, New York [u.a.], 2. ed edition, 1991.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM.

[6] K. Emamy and R. Cameron. Citeulike: A researcher's social bookmarking service. `http://www.ariadne.ac.uk/issue51/emamy-cameron/`, 2007.

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

[8] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *CoRR—Computing Research Repository*, abs/cs/0508082, 2005.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, volume 4011 of *LNCS*, pages 411–426, Budva, Montenegro, June 2006. Springer.

[10] V. A. Koutsonikola, A. Vakali, E. Giannakidou, and I. Kompatsiaris. Clustering of social tagging system users: A topic and time based approach. In G. Vossen, D. D. E. Long, and J. X. Yu, editors, *WISE'09 : Proceeding of the 10th International Conference Web Information Systems Engineering*, volume 5802 of *Lecture Notes in Computer Science*, pages 75–86. Springer, 2009.

[11] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[12] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. X. Parreira, M. Spaniol, and G. Weikum. Social wisdom for search and recommendation. *IEEE Data Eng. Bull.*, 31(2):40–49, 2008.

[13] S. Sen, J. Vig, and J. Riedl. Tagommenders: connecting users to items through tags. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *WWW*, pages 671–680. ACM, 2009.

[14] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing and Management*, 32(5):619–633, 1996.

[15] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 515–522. ACM, 2008.

[16] N. Zhang, Y. Zhang, and J. Tang. A tag recommendation system for folksonomy. In *SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining*, pages 9–16, New York, NY, USA, 2009. ACM.