

Analyse du comportement d'annotation du réseau social d'un utilisateur pour la détection des intérêts - Application sur Delicious

Manel Mezghani, Corinne Zayani, André Péninou, Ikram Amous, Florence Sèdes

▶ **To cite this version:**

Manel Mezghani, Corinne Zayani, André Péninou, Ikram Amous, Florence Sèdes. Analyse du comportement d'annotation du réseau social d'un utilisateur pour la détection des intérêts - Application sur Delicious. Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information, 2015, vol. 4, pp. 85-111. 10.3166/isi.20.4.85-111 . hal-01371778

HAL Id: hal-01371778
https://hal.science/hal-01371778

Submitted on 26 Sep 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 : 15323

To link to this article : DOI : 10.3166/isi.20.4.85-111
URL : http://dx.doi.org/10.3166/isi.20.4.85-111

To cite this version : Mezghani, Manel and Péninou, André and Zayani, Corinne Amel and Amous, Ikram and Sèdes, Florence *Analyse du comportement d'annotation du réseau social d'un utilisateur pour la détection des intérêts - Application sur Delicious.* (2015) Ingénierie des Systèmes d'Information, vol. 4. pp. 85-111. ISSN 1633-1311

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

Analyse du comportement d'annotation du réseau social d'un utilisateur pour la détection des intérêts

Application sur *Delicious*

Manel Mezghani[1,2], **André Péninou**[2], **Corinne Amel Zayani**[1], **Ikram Amous**[1], **Florence Sèdes**[2]

1. Laboratoire MIRACL, Université de Sfax
3021 SFAX, Tunisie
mezghani.manel@gmail.com
{corinne.zayani, ikram.amous}@isecs.rnu.tn

2. Institut de Recherche en Informatique de Toulouse (IRIT)
Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J
31062 Toulouse Cedex 9, France
{manel.mezghanni, andre.peninou, florence.sedes}@irit.fr

RÉSUMÉ. *L'utilisateur social est caractérisé par son activité sociale comme le partage d'informations et l'établissement de relations avec d'autres utilisateurs. Avec l'évolution du contenu social, l'utilisateur a besoin d'informations plus précises qui reflètent ses intérêts. Nous nous concentrons sur la détection des intérêts de l'utilisateur qui sont des éléments clés pour améliorer l'adaptation (recommandation, personnalisation, etc.). L'originalité de notre approche est basée sur la proposition d'une nouvelle technique de détection des intérêts qui analyse le réseau des relations d'un utilisateur et aussi la précision de leurs comportements d'annotation dans le but de sélectionner les tags qui reflètent réellement le contenu des ressources. L'approche proposée a été testée et évaluée sur la base de données sociales Delicious. Pour l'évaluation, nous comparons le résultat issu de notre approche utilisant le comportement d'annotation des personnes proches (le réseau égocentrique ou les communautés) avec les informations connues de l'utilisateur (son profil). Une évaluation comparative avec une approche classique (basée sur les tags) de détection des intérêts montre que l'approche proposée fournit de meilleurs résultats.*

ABSTRACT. *The social user is characterized by his social activity such as sharing information and making relationships. With the evolution of social content, the user needs more accurate information that reflects his interests. We focus on detecting user's interests which are key elements for improving adaptation (recommendation, personalization, etc.). The originality of our approach is based on the proposal of a new technique of interests detection by analysing the*

accuracy of the tagging behaviour of the social network of a user in order to figure out the tags which really reflect the content of the resources. The proposed approach has been tested and evaluated in the Delicious social database. For the evaluation, we compare the result issued from our approach using the tagging behaviour of the neighbours (the egocentric network or the communities) with the information yet known for the user (his profile). A comparative evaluation with the classical tag-based method of interests detection shows that the proposed approach is better.

MOTS-CLÉS : *profil utilisateur, intérêts, réseaux sociaux, indexation, comportement d'annotation.*

KEYWORDS: *user profile, interests, social network, indexation, tagging behaviour.*

1. Introduction

L'avènement du web 2.0, centré utilisateur, a fait émerger une quantité très importante d'informations. Souvent partagées dans les médias sociaux, ces informations constituent un moyen pour guider les autres utilisateurs vers l'information recherchée. Cet aspect collaboratif de partage d'information est utilisé dans plusieurs applications comme le e-commerce, le e-learning, etc. Cependant, cette quantité d'informations rend leur accès de plus en plus difficile compte tenu de la diversité du contenu qui peut intéresser l'utilisateur. Plusieurs techniques ont été développées afin de mieux utiliser la connaissance collective partagée par les utilisateurs du réseau social. Parmi ces techniques, citons l'adaptation qui permet de fournir à l'utilisateur une information qui convient mieux à répondre à ses besoins. L'adaptation est un terme assez générique et se décline sous plusieurs formes : personnalisation, recommandation, etc.

L'adaptation est donc un processus fortement lié à l'utilisateur puisqu'on adapte l'information selon les besoins spécifiques de celui-ci. Un profil utilisateur [1] qui reflète les intérêts réels de l'utilisateur permet d'améliorer l'adaptation et ainsi d'éviter la surcharge cognitive et la désorientation de l'utilisateur pendant son accès à l'espace d'information.

Il existe de nombreuses sources pour extraire les intérêts d'un utilisateur. Le profil de l'utilisateur contient des informations fournies explicitement par l'utilisateur lui-même, par exemple : nom, prénom, âge, profession, etc. De manière classique, les intérêts d'un utilisateur sont extraits de son propre profil (par exemple de l'attribut intérêt). Les intérêts peuvent être également extraits de son comportement social ou par exemple le comportement d'annotation, ou de son réseau social ou par exemple ses amis. Toutefois, dans un contexte social, de nombreux paramètres font de la détection d'intérêts un problème complexe. Nous avons choisi de nous concentrer sur certains problèmes qui affectent le processus de détection des intérêts :

1. Nous considérons dans cet article qu'un profil utilisateur est un vecteur de mots-clés (intérêts).

1. **Le manque d'information fournie par l'utilisateur lui-même** : L'utilisateur ne donne pas généralement de façon exhaustive les informations relatives à ses intérêts. Donc, le profil explicite de l'utilisateur ne peut jamais être considéré comme entièrement connu par le système. Ainsi, il est difficile de s'appuyer sur la seule analyse du profil explicite pour détecter les intérêts pertinents (Tchuente *et al*., 2013).

2. **L'activité dense de l'utilisateur social** : L'utilisateur est de plus en plus actif : il participe à des discussions, commente et annote des ressources, etc. Par conséquent, détecter ses intérêts devient plus difficile (Ma *et al*., 2011). En effet, différents comportements de l'utilisateur peuvent décrire ses intérêts ce qui rend ensuite le choix du comportement à analyser difficile.

3. **La variété et la quantité des ressources** : Une approche sociale de détection d'intérêts doit faire face à l'aspect évolutif des médias sociaux. La quantité d'informations (contenu et utilisateurs) est en croissance exponentielle. Pour les utilisateurs, de nombreuses relations peuvent être établies (relations d'amitié, utilisateurs appartenant au même groupe, etc.). Pour le contenu, de nombreux types d'informations sont disponibles dans les médias sociaux tels que des images, pages web, vidéos, etc. Cette variété rend la détection d'intérêts plus difficile, puisque l'utilisateur peut interagir avec plusieurs contenus.

4. **La qualité potentiellement mauvaise des annotations (tags)** : En effet, les tags reflètent l'opinion d'un utilisateur vis-à-vis d'une ressource. Mais, ils sont générés par l'utilisateur (mots libres) et peuvent ainsi être non compréhensibles. Cette caractéristique peut induire une mauvaise compréhension des tags par le système ou même par les autres utilisateurs.

Notre approche est construite à partir de l'hypothèse que l'environnement social et en particulier les personnes proches d'un individu fournissent une information à partir de laquelle les intérêts réels de cet individu peuvent être extraits. Cette hypothèse a été prouvée dans le contexte de dérivation des profils d'utilisateurs à partir de leurs réseaux sociaux (Tchuente *et al*., 2013). Nous analysons les personnes proches afin de détecter les intérêts les plus pertinents pour chaque utilisateur. Les personnes proches peuvent être celles partageant certains comportements communs (par exemple visitant ou annotant la même ressource), le réseau égocentrique [2], les utilisateurs appartenant à la même communauté [3], etc. Nous étudions plusieurs cas possibles.

Nous nous concentrons sur le comportement d'annotation des personnes proches, qui reflète l'opinion de ces utilisateurs sur une ressource (Astrain *et al*., 2010). Cette information a prouvé son utilité pour détecter les intérêts des utilisateurs (Meo *et al*., 2010) (Kim *et al*., 2011). Elle est représentée sous forme de tuples <utilisateur, tag, ressource> dénotant qu'un utilisateur a posé un tag sur une ressource.

---

2. L'ensemble des individus avec qui l'utilisateur est en relation explicite directe.

3. La communauté est définie par un ensemble d'utilisateur, qui partagent des caractéristiques en commun. Ces dernières varient d'un contexte à un autre.

L'approche proposée traite principalement les ressources textuelles (semi-structurées, texte, etc.) qui sont présentes dans presque tous les médias sociaux tels que : *Delicious* en analysant l'URL annotée, *Twitter* en analysant les *tweets*, etc. Notre approche ne traite pas les autres médias (les images dans le cas de *Flickr* par exemple).

Notre approche est expérimentée sur la base de données sociales *Delicious*. Différents types de personnes proches sont considérées dans notre expérimentation : i) le réseau égocentrique, ii) les utilisateurs appartenant à la même communauté. Nos résultats sont comparés à l'approche construisant un profil à partir des tags fournis par les utilisateurs (approche classique basée sur les tags sans tenir en compte des ressources auxquelles ils sont associés). Cette approche est la continuation des travaux menés dans (Mezghani *et al*., 2014) et (Mezghani *et al*., 2015).

Le reste de cet article est structuré comme suit. Dans la deuxième section, nous présentons un état de l'art sur la détection des intérêts à partir du comportement social. Dans la troisième section, nous présentons notre positionnement par rapport à l'état de l'art comme une combinaison d'approches. Dans la quatrième section, nous présentons et décrivons l'approche proposée qui s'appuie principalement sur un filtrage des tags à partir des ressources et un réseau de relations de l'utilisateur. Dans la cinquième section, nous présentons et commentons les résultats de notre expérimentation sur la base sociale *Delicious*, qui donne de meilleurs résultats que les approches classiques. Dans la dernière section, nous concluons et présentons les perspectives de notre travail.

2. Approches de détection des intérêts à partir du comportement social

Selon (Astrain *et al*., 2010), les intérêts peuvent être déduits de l'environnement social en se basant sur trois sources d'information : **l'utilisateur**, **la ressource** et/ou **le tag**. Nous présentons quelques recherches se concentrant sur chaque élément.

2.1. Détection des intérêts basée sur les utilisateurs (les personnes proches)

Les personnes proches sont considérées comme une source importante d'information pour la détection des intérêts. En effet, les informations issues des personnes proches ont prouvé leur utilité pour surmonter le problème du démarrage à froid (*cold-start*) pour les nouveaux utilisateurs du système, pour détecter les intérêts des utilisateurs (Tchuente *et al*., 2013) (Canut *et al*., 2015) et aussi pour enrichir les profils utilisateurs dans un but de recommandation (Meo *et al*., 2010), (Kim *et al*., 2011). La définition d'une personne proche est la relation sociale de l'utilisateur avec d'autres utilisateurs. Cette relation peut être explicite (une relation d'amitié) ou implicite (les utilisateurs qui agissent sur le même document par exemple). Ces relations sociales sont détaillées dans (Musiał, Kazienko, 2013). Les personnes proches de l'utilisateur dans un contexte social, sont décrites par des liens, où un lien entre deux utilisateurs agrège tous les types des relations qui existent entre ces deux personnes (Musiał, Kazienko, 2013).

Une personne proche peut être une « bonne » personne qui influence l'utilisateur d'une manière positive. Ceci servira pour dériver ou enrichir le profil utilisateur selon l'information présente dans les profils de ses personnes proches, pour recommander les ressources pertinentes selon les ressources visitées par ses personnes proches, etc. Si c'est une « bonne » personne alors le résultat obtenu sera jugé profitable par l'utilisateur. Une personne proche peut être aussi une « mauvaise » personne qui influence l'utilisateur d'une manière négative comme les « spammeurs » dont le but est d'inonder le système d'information visant à désorienter l'utilisateur.

Des études analysent l'environnement social afin de détecter les personnes proches (en terme d'intérêts). Ces personnes proches sont détectées par plusieurs métriques telles que la similarité cosinus (Kim *et al*., 2011), « X-compass » (Meo *et al*., 2010), etc.

D'autres études détectent les personnes proches par des observations, comme (Kim *et al*., 2011) qui enrichissent le profil d'un utilisateur avec des tags de ses amis non inclus dans le profil. Ils se basent sur l'observation que deux personnes sont proches si elles partagent des tags communs et donc elles peuvent bien avoir des intérêts en commun.

D'autres chercheurs visent à combiner différents paramètres afin de détecter la proximité entre les personnes. (Cabanac, 2011) calcule la proximité entre les auteurs scientifiques par une combinaison de leur distance (le degré de séparation dans le graphe des co-auteurs), leur connectivité (le nombre des chemins dans le graphe entre deux auteurs) et le nombre de papiers écrits en commun. Dans le cadre de la recommandation sociale, (Guy *et al*., 2010) calculent le score de la proximité à partir de différents critères : i) le nombre de personnes et/ou de tags dans le profil utilisateur qui sont liés à un article, ii) le degré de connectivité de ces personnes et/ou des tags à un utilisateur, iii) le degré de connectivité de ces personnes et/ou tags à un article, et iv) la fraîcheur d'un article. (Roth *et al*., 2010) détectent les relations implicites entre les utilisateurs par leur échange de courrier. Ils calculent la proximité par la fréquence de l'interaction entre les utilisateurs, la fraîcheur de l'interaction et la direction de l'interaction.

2.2. Détection des intérêts basée sur les ressources

Les intérêts peuvent être déduits sur la base des ressources que l'utilisateur accède (Ma *et al*., 2011) (White *et al*., 2009). La ressource peut être de n'importe quel type (URL, vidéo, image, etc.). Dans (Ma *et al*., 2011) les intérêts des utilisateurs sont découverts par extraction et par analyse des mots-clés de chaque source (les sources sont *Facebook, LinkedIn,* etc.). Dans (White *et al*., 2009) les intérêts des utilisateurs sont découverts à partir de l'analyse du comportement de l'utilisateur c'est-à-dire l'historique de visite des ressources, le temps passé sur une page web, etc.

Pour analyser le contenu de la ressource, différentes techniques existent comme l'indexation qui est utilisée pour extraire les termes significatifs des ressources. Après

l'indexation des ressources, différentes fonctions de score peuvent être appliquées afin de détecter la ressource la plus pertinente selon une requête spécifique (Vallet *et al*., 2010). Ces fonctions, appliquées dans un contexte de recherche d'information, peuvent être le TF * IDF, BM25, etc. (Vallet *et al*., 2010). Ces scores sont le résultat d'un processus d'indexation, qui invoque une requête et une collection de ressources. L'utilisation de ces méthodes a montré leur utilité et robustesse dans la recherche d'information (Cai, Li, 2010).

Dans un contexte social, la requête peut être un tag et le contenu des ressources annotées a été analysé dans un but de recommandation dans une perspective d'apprentissage automatique dans (Song *et al*., 2011). En outre, (Zhang *et al*., 2010) proposent une approche de recommandation (modélisée comme une approche basée sur *Latent Dirichlet Allocation*) dans les systèmes à base de tags. Cette approche combine le contenu et l'analyse des relations dans un modèle unique.

2.3. Détection des intérêts basée sur les tags

Plusieurs recherches portent sur la détection des intérêts de l'utilisateur à partir de l'information sociale produite par les utilisateurs et particulièrement des tags. Ces derniers sont considérés comme une information puissante pour refléter l'opinion de l'utilisateur vis-à-vis d'une ressource (Meo *et al*., 2010) et aussi pour détecter les intérêts de l'utilisateur (Kim *et al*., 2011).

Les travaux qui détectent des intérêts en analysant les annotations sociales (tags), analysent : les tags les plus récents (Zheng, Li, 2011), les tags plus populaires (Godoy, Amandi, 2008), l'historique des annotations (Wang *et al*., 2010), ou la sémantique des tags (Kim *et al*., 2011).

Cependant, les tags sont des mots-clés générés par l'utilisateur et donc ils ne suivent aucune règle. Par conséquent, ils peuvent contenir une information ambiguë et qui ne reflète pas le contenu de la ressource. Par exemple un tag peut être : i) un spam (qui vise à promouvoir un intérêt d'un autre utilisateur par exemple) ou ii) un tag personnel (qui reflète le « sentiment » de l'utilisateur et non pas le contenu de la ressource comme : bien, j'aime, nul, etc.) ou bien iii) un mot propre à un utilisateur et qui n'est pas compréhensible soit par les autres utilisateurs, soit par le système.

L'ensemble des tags affecté aux ressources est appelé *folksonomie* (Meo *et al*., 2010), (Milicevic *et al*., 2010) et (Vallet *et al*., 2010). La *folksonomie* est un outil puissant pour capturer la connaissance collective (Milicevic *et al*., 2010). Contrairement à l'ontologie, la *folksonomie* n'est pas structurée. Cette caractéristique conduit à un vocabulaire ambigu qui peut influencer la mauvaise compréhension des intérêts des utilisateurs par le système ou même par d'autres personnes dans le réseau. Pour éviter ces problèmes, de nombreuses mesures sont proposées comme : *SpamFactor, SpamRank, spamClean*, etc. (Milicevic *et al*., 2010). En outre, plusieurs techniques sont utilisées telles que le *clustering*, la conversion de la *folksonomie* en une ontologie (Carmagnola *et al*., 2007) ou d'une manière classique en utilisant un outil de

traitement du langage naturel comme *WordNet* [4]. Ces techniques visent à structurer la *folksonomie* d'une manière compréhensible pour l'utiliser dans un but de recommandation, de personnalisation, etc. Plus de détails sur le traitement de l'ambiguïté des tags sont présentés dans (Milicevic *et al*., 2010).

3. Positionnement

Un système d'adaptation sociale efficace essaye de détecter les intérêts des utilisateurs à l'aide de données sociales pertinentes. Mais les intérêts estimés d'un utilisateur peuvent être considérés comme non pertinents en raison de l'information inappropriée utilisée pour les détecter. Pour surmonter ce problème, notre approche cherche à réaliser une utilisation sélective de l'information disponible pour produire une liste d'intérêts précise pour chaque utilisateur.

De l'information disponible, nous choisissons l'information du comportement d'annotation (cf. section 2).

Nous détaillons notre analyse ci-dessous :

– **Utilisateur** : Afin de développer notre approche, nous analysons le comportement d'annotation des personnes proches de chaque utilisateur. Nous analysons ainsi le réseau de relations (réseau égocentrique ou communautés). Ce choix est motivé par :

a) les études qui favorisent la connaissance collective pour détecter les intérêts des utilisateurs (Meo *et al*., 2010), (Kim *et al*., 2011), (Tchuente *et al*., 2013) et (Zhou *et al*., 2010).

b) l'absence d'informations dans le profil d'utilisateur explicite. En effet, l'utilisateur ne fournit pas toutes les informations relatives à ses intérêts. Donc son profil ne peut jamais être considéré comme une information suffisante pour connaître ses intérêts (Tchuente *et al*., 2013).

c) l'information inappropriée déduite uniquement à partir de son comportement classique. En fait, le comportement de l'utilisateur ne reflète pas toujours les vrais intérêts de l'utilisateur. Par exemple, l'analyse du comportement de navigation de l'utilisateur selon (Ma *et al*., 2011) : i) conduit à l'analyse du comportement antérieur qui peut ne pas refléter ses intérêts actuels, ii) n'est pas toujours un indicateur efficace puisque l'utilisateur peut accéder à une page web sans avoir un intérêt pour son contenu.

– **Tag** : Dans la plupart des travaux analysant le comportement d'annotation, les intérêts sont détectés à partir des tags. Cette détection est basée sur des mesures de popularité des tags ou d'analyse des tags (par analyse de la sémantique des tags, par exemple) (Milicevic *et al*., 2010). Ces analyses peuvent fournir des tags pertinents pour l'utilisateur. Mais, selon (Milicevic *et al*., 2010), le problème lié à certains tags est qu'ils sont spécifiques à l'utilisateur. En effet, ces tags ne décrivent pas le do-

4. http://wordnet.princeton.edu/

cument mais plutôt l'avis de l'utilisateur comme par exemple « j'aime » ,« sympa » ,« nul », etc. Selon (Milicevic *et al*., 2010), l'ambiguïté d'un tag est qu'un seul tag a de nombreuses significations et peut faussement donner l'impression que deux ressources sont similaires quand elles sont en fait sans rapport. Ainsi, le filtrage des tags pourrait être une solution pour surmonter les tags ambigus. Donc, dans un but de détecter des intérêts pertinents, nous allons essayer de produire des tags significatifs (compréhensibles) plutôt que les tags spécifiques à un utilisateur.

– **Ressource** : Généralement, les approches traitant cette information utilisent des techniques telles que le *clustering*, analyse sémantique, etc. (Ma *et al*., 2011). Cependant, n'analyser que le contenu des ressources accédées ne permet pas d'avoir des information sur les intérêts de l'utilisateur (Ma *et al*., 2011). Néanmoins, le contenu des ressources peut se révéler capable de détecter la nature des tags qui lui sont associés. La plupart des recherches ne considèrent pas l'exactitude des tags avec le contenu de la ressource (Zhou *et al*., 2010). Contrairement à la plupart de ces recherches, nous nous concentrons sur l'analyse de l'exactitude des tags par rapport au contenu des ressources pour surmonter les problèmes liés à la nature des annotations sociales.

Pour résumer, notre approche tente de combiner les informations utilisateur, tag et ressource d'une manière qui cherche à garantir une détection des intérêts pertinents. Afin d'obtenir des intérêts (tags) pertinents : i)les tags retenus sont ceux qui reflètent le contenu des ressources auxquelles ils ont été associés, ii) les intérêts sont détectés à partir de personnes proches.

4. Description de l'approche de détection des intérêts

L'approche de détection des intérêts que nous proposons analyse, d'une part le réseau de relations (réseau égocentrique ou communautés) d'un utilisateur, et, d'autre part, la précision du comportement d'annotation des utilisateurs du réseau de relations dans le but de sélectionner les tags qui reflètent réellement le contenu des ressources. Le filtrage de ces tags se fait en plusieurs étapes :

1. Pour chaque tag de chaque utilisateur, construction de l'ensemble des ressources pertinentes pour ce tag. Cette construction est effectuée par l'analyse de l'ensemble de toutes les ressources pour un tag donné,

2. Pour chaque tag, attribution d'un score à ces ressources et sélection des top-k ressources,

3. Filtrage des tags : si une ressource associée au tag est dans le top-k des ressources pertinentes pour ce tag, alors le tag est retenu.

Le fait d'utiliser toutes les ressources existantes pour chercher les ressources pertinentes pour un tag (et non pas les seules ressources auxquelles ce tag est associé) doit permettre de réellement analyser la pertinence des tags. Un tag ne sera retenu que s'il est associé à une ressource qui appartient à l'ensemble des ressources auxquelles il correspond le mieux (calcul des top-k ressources pour chaque tag). Nous cherchons ainsi à analyser la précision du comportement d'annotation d'un utilisateur dans le but

de sélectionner les tags qui reflètent réellement le contenu des ressources et reflètent alors le mieux les intérêts de l'utilisateur.

L'approche de détection des intérêts est effectuée selon deux étapes principales détaillées ci-après. D'abord, nous préparons les données que nous allons utiliser. Ensuite, nous détaillons le processus de détection des intérêts (filtrage des tags).

Pour la suite de l'article, notons :

– $U = \{u_1, \ldots, u_n\}$, l'ensemble des utilisateurs dans le réseau social, où n est le nombre d'utilisateurs.

– $R = \{r_1, \ldots, r_m\}$, l'ensemble de toutes les ressources dans le réseau social, où m est le nombre de ressources.

– $T = \{t_1, \ldots, t_h\}$, l'ensemble des tags, où h est le nombre de tags.

– $N_u = \{n_{u1}, \ldots, n_{uj}\}$, l'ensemble des personnes proches de l'utilisateur u, où j est le nombre de personnes proches de l'utilisateur $u \in U$.

– $I_u = \{i_{u1}, \ldots, i_{uk}\}$, l'ensemble des intérêts pertinents pour l'utilisateur u, où k est le nombre des intérêts pertinents de l'utilisateur $u \in U$. Ceci est le résultat construit par notre algorithme.

4.1. Préparation des données

Avant d'expliquer notre processus de détection des intérêts, nous procédons à la préparation des données utilisées comme une entrée pour détecter les intérêts des utilisateurs. La figure 1 illustre cette étape de préparation de données.

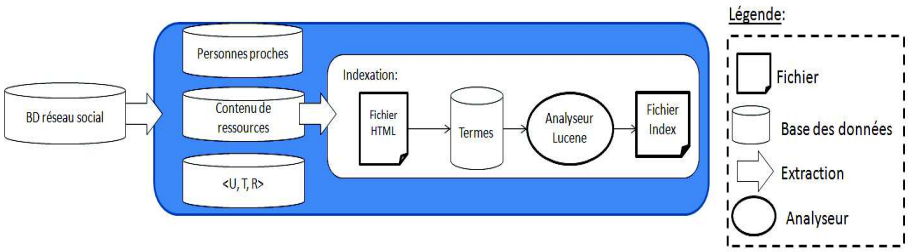


Figure 1. Préparation des données

Nous extrayons les données relatives au :

– Comportement d'annotation $< U, T, R >$, constitués par les tags T appliqués aux ressources R par les utilisateurs U,

– Personnes proches N_u (le réseau égocentrique ou les communautés),

– Contenu des ressources.

Après avoir extrait les données, nous indexons les ressources extraites. L'indexation vise à décrire le contenu d'un document par des mots clés. Les ressources sont indexées (comme les ressources semi-structurées ou texte brut), en utilisant *Lucene*

API[5]. *Lucene* est capable de traiter de grands volumes de documents grâce à sa puissance et à sa rapidité dues à l'indexation. *Lucene* est un outil d'indexation basé sur les champs. Cette caractéristique permet l'indexation des ressources selon un ou plusieurs champs. Par exemple, les champs peuvent être le *titre*, le *contenu*, l'*URL*, etc. Nous n'avons tenu en compte que du champ *contenu*, vu la richesse de l'information présente.
Les étapes de cette technique d'indexation sont comme suit : *Lucene* indexe les ressources avec un parseur[6] en les divisant en un certain nombre de termes en utilisant un analyseur. Puis, il stocke les termes dans un fichier d'index, où chaque terme est associé avec le contenu de la ressource.

L'index est composé de segments, pouvant être considérés comme des sous-index bien qu'ils ne soient pas entièrement indépendants. *Lucene* va assigner à chaque document de l'index un identifiant unique (Document ID). Les segments conservent les éléments suivants : 1) les noms des champs utilisés dans l'index, 2) un dictionnaire des termes : les termes contenus dans chaque champ, 3) la fréquence des termes : numéros de tous les documents contenant ce terme et 4) proximité des termes : la position de chaque terme.

4.2. Processus de détection des intérêts à partir du comportement social

Nous présentons l'algorithme général de notre processus de détection des intérêts du profil utilisateur dans le tableau 1. Cet algorithme est appliqué pour tous les utilisateurs U.

Entrée : N_u, T, *FichierIndex*;
Sortie : I_u;
1: Début
2: **Pour** chaque $n_{uj} \in N_u$ **faire**
3: **Pour** chaque $t_h \in T$ **faire**
4: R'= GenerationRessourcesPertinentsAuTag(t_h, *FichierIndex*);
5: R"=Top-k Score(R',t_h);
6: **Si** $\exists\, r_{v''} \in R''$, $t_h \exists < U, T, R'' >$ **alors**
7: I_u=I_u+ t_h;
8: **Fin Si**
9: **Fin Pour**
10: **Fin Pour**
11: **Retourner** I_u;
12: Fin

Algorithme 1 : Algorithme général du processus de détection des intérêts pour un utlisateur *u*

5. http://lucene.apache.org/core/
6. Analyseur syntaxique qui étiquette les mots d'un texte.

Cet algorithme procède pour chaque personne proche ($n_{uj} \in N_u$) d'un utilisateur donné ($u \in U$) comme suit : d'abord, il génère les ressources pertinentes (R') pour chaque tag ($t_h \in T$). Cette génération a pour but de sélectionner les tags qui reflètent le contenu de la ressource. Puis, il score ces ressources afin de ne conserver que les ressources les plus pertinentes (R"). Enfin, il sélectionne les tags pertinents qui sont associés à la fois à la ressource annotée par u et à la liste des ressources pertinentes (R"). Ces tags sont considérés comme étant des intérêts de l'utilisateur (I_u) puisqu'ils reflètent vraiment le contenu de la ressource. Ce processus est illustré dans la figure 2.

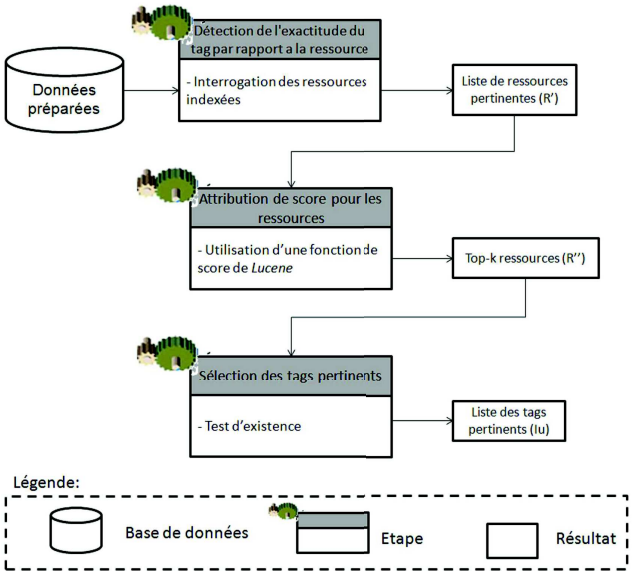


Figure 2. Processus de détection des intérêts pour un utilisateur pour chacune de ses personnes proches

Afin d'expliquer notre approche, nous montrons un exemple de détection d'un intérêt à l'aide de l'analyse d'un tag d'une personne proche. Nous montrons dans la figure 3 l'exemple d'un réseau social très simplifié d'un utilisateur.

A partir de chaque personne proche de la figure 3, nous analysons les tags utilisés. Nous montrons dans la figure 4 un exemple de détection d'un intérêt à l'aide de l'analyse d'un tag d'une personne proche. Cet exemple est réitéré pour chaque tag de chaque personne proche afin de construire toute la liste des intérêts pertinents pour l'utilisateur.

En prenant le tag *t8* associé à la ressource *r8* par la personne proche *u3*, l'algorithme commence par générer la liste des ressources pertinentes *R'*. Cette liste est le résultat de l'interrogation du fichier Index. La deuxième étape consiste à prendre seulement les top-k ressources de *R'* (dans la figure nous avons pris le top-2 ressources) afin de construire *R"*. La troisième étape consiste à tester si la ressource *r8*

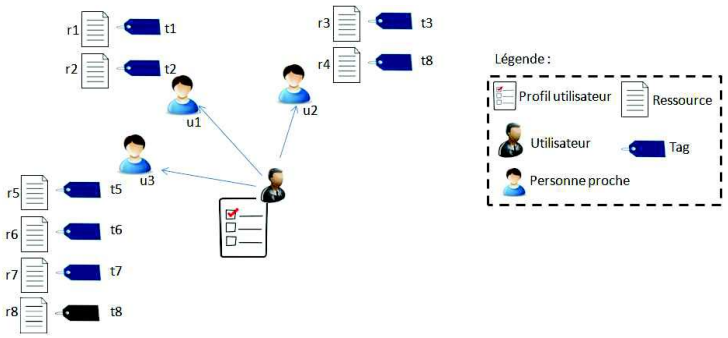


Figure 3. L'exemple d'un réseau social d'un utilisateur

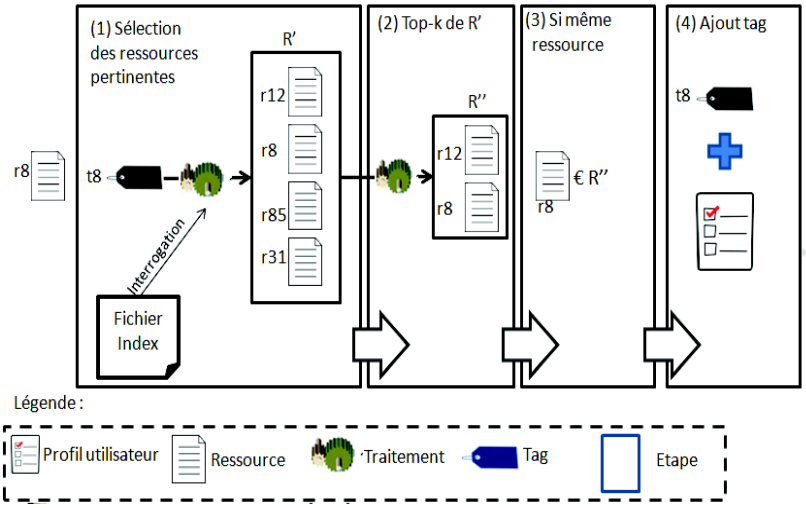


Figure 4. Un exemple de détection d'un intérêt à travers l'analyse d'un tag d'une personne proche

(associé au tag *t8*) appartient à *R''*. Comme c'est le cas, *t8* sera jugé comme un intérêt potentiel pour l'utilisateur.

Dans la figure 3, nous remarquons que le tag *t8* est aussi associé à la ressource *r4* par l'utilisateur *u2*. Lors du traitement de ce triplet <u2, r4, t8>, *r4* n'appartenant pas à *R''*, le tag t8 ne sera pas retenu dans cette étape comme étant un intérêt potentiel pour l'utilisateur. Néanmoins, comme précisé dans le paragraphe précédent, *t8* sera bien sûr retenu comme intérêt potentiel lors du traitement du triplet <u3, r8, t8>.

4.2.1. Détection de l'exactitude du tag par rapport à la ressource

Nous commençons par générer les ressources pertinentes R' pour chaque tag donné, où $R' = \{r'_1, \ldots, r'_v\}$ est l'ensemble des ressources pertinentes et *v* est le nombre de

ressources pertinentes et $R' \subseteq R$. Cette étape développe la fonction *GenerationRessourcesPertinentsAuTag(t_h, FichierIndex)* de l'algorithme générale (cf. algorithme 1). Cette étape interroge le *Fichier Index* (la sortie de l'étape d'indexation). Lorsqu'une requête est faite, elle est traitée par le même analyseur utilisé pour créer l'index et ensuite utilisé pour trouver le(s) expression(s) correspondante(s) dans l'index. Ceci fournit une liste de ressources correspondant à la requête. Dans notre contexte, une requête est considérée comme un tag dans le reste de ce papier.

4.2.2. Attribution de score pour les ressources

Après la génération des ressources pertinentes (R') pour chaque tag (t_h), un score est attribué à chaque ressource pertinente. Le but de l'utilisation d'un tel score est de sélectionner les ressources les plus pertinentes liées à un tag. Ce score est le résultat d'une fonction de similarité qui prend en considération la ressource et le tag. De nombreuses fonctions de similarité existent dans la littérature telles que la fonction de similarité soutenue par *Lucene*. Nous choisissons une fonction prédéfinie [7] de similarité qui est une variante du modèle de notation TF-IDF. Le choix d'un tel modèle est dû au fait que TF-IDF est un algorithme simple et efficace pour faire correspondre des mots dans une requête aux ressources qui sont pertinentes pour cette requête.

Cette fonction fournit les top-k ressources pertinents (R") pour un tag, où R"= {r"$_1$,..., r"$_w$} est l'ensemble des top-k ressources pertinentes et w est le nombre de ressources pertinentes et $R'' \subseteq R'$. Cette étape développe la fonction *Top-k Score (R',t_h)* de l'algorithme générale (cf. algorithme 1).

4.2.3. Sélection des tags pertinents

Après avoir généré les top-k ressources pertinentes pour un tag t_h (d'une personne proche), nous testons si la ressource annotée par t_h (la ressource annotée directement par l'utilisateur) existe dans le résultat top-k fournie par la fonction de score. Si c'est le cas, le tag t_h est considéré comme pertinent pour la ressource (puisqu'il reflète vraiment son contenu).
Cette étape est représentée par les lignes 6, 7 et 8 de l'algorithme 1. Elle génère une liste des intérêts pertinents (I_u) comme une liste de tags qui décrivent au mieux le contenu de la ressource annotée. Cette liste est issue de l'analyse des personnes proches pour chaque utilisateur.

4.3. Processus de validation

Afin de valider notre approche, nous considérons les utilisateurs ayant un profil connu (i.e. qui ont déjà eu des activités dans le réseau). Dans une approche classique, nous considérons le profil individuel de l'utilisateur comme étant la liste de tags affectés par l'utilisateur sur des ressources. Donc, nous comparons les tags de l'utilisateur

7. http://lucene.apache.org/core/3_ 5_ 0/scoring.html

(issus de son profil) avec des tags fournis par notre approche (issus des personnes proches : réseau égocentrique ou communautés).

De notre analyse sociale nous avons construit, pour chaque utilisateur, une liste d'intérêts (tags). Cette liste est validée par sa comparaison avec les intérêts de cet utilisateur. Le processus d'évaluation est décrit dans la figure 5.

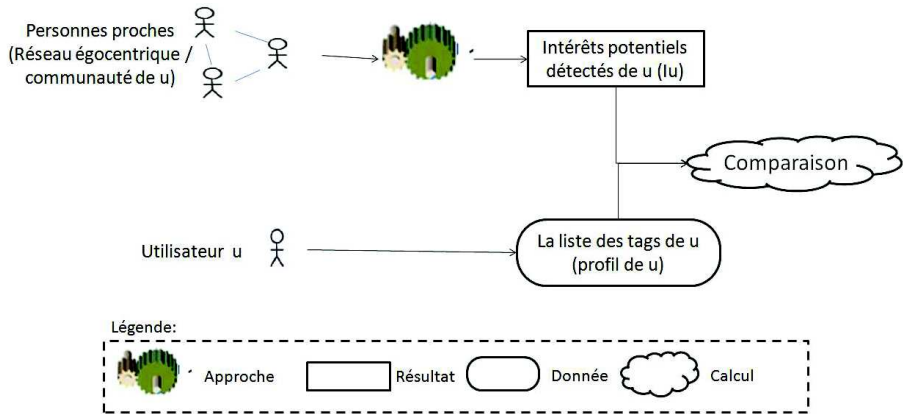


Figure 5. Processus d'évaluation de l'approche de détection des intérêts

Pour chaque utilisateur cible $u \in U$, notre approche construit un ensemble I_u= $\{i_{u1}, \ldots, i_{uk}\}$ d'intérêts potentiellement pertinents. Pour chaque $i_{uk} \in I_u$, nous analysons l'existence de l'intérêt i_{uk} dans le profil de l'utilisateur cible u. Les intérêts corrects (réellement pertinents) sont représentés par l'ensemble des intérêts $C_u \subset I_u$, où C_u= $\{c_{u1}, \ldots, c_{uy}\}$ et y est le nombre d'intérêts présents à la fois dans I_u et le profil individuel de *u*.

La validation de notre proposition se fait par un test d'existence des intérêts réels des utilisateurs (profil individuel) dans les intérêts potentiels calculés par notre approche. Ce test est effectué à l'aide de deux méthodes :

– Par une simple comparaison des tags (comparaison exacte) : par exemple si un tag de l'utilisateur= « image » et un tag trouvé par notre approche= « image », alors « image » est considéré comme un tag pertinent. Nous appellerons cette technique dans le reste de cet article « simple comparaison »,

– En prenant en compte les synonymes ou les mots reliés : par exemple si un tag de l'utilisateur= « image » et un tag trouvé par notre approche= « photo », alors « image » est considéré comme un tag pertinent. Les synonymes ou mots reliés sont détectés en interrogeant Wordnet [8]. Nous appellerons cette technique dans le reste de cet article « avec synonymes ou mots reliés ».

8. http://wordnet.princeton.edu/

5. Expérimentations sur *Delicious*

Dans cette section, nous détaillons d'abord la base de test utilisée dans la section 5.1. Ensuite, nous présentons les mesures utilisées pour nos calculs dans la section 5.2. Enfin, nous détaillons les évaluations faites afin de tester l'efficacité de notre approche. En effet, notre approche est évaluée selon deux critères.

Tout d'abord, nous étudions dans la section 5.3 l'influence de l'environnement social de l'utilisateur, et principalement l'influence des personnes proches dans la précision des résultats. Nous avons évalué notre approche de deux façons : en utilisant le réseau égocentrique de l'utilisateur et en utilisant les communautés de l'utilisateur (généré à partir d'un algorithme spécifique de détection de communautés). Nous avons aussi évalué notre approche selon les deux méthodes de validation : i) par une technique de comparaison simple ou ii) en tenant compte des synonymes ou mots reliés. Nous avons également testé l'influence de la valeur de k qui sélectionne les top-k ressources pertinentes pour un tag. Nous conservons les valeurs qui fournissent de meilleurs résultats pour faire le reste des évaluations.

Puis, nous comparons dans la section 5.4 notre approche avec l'approche qui utilise les informations des tags des personnes proches sans pré-traitement (approche classique basée sur les tags [9]). Pour être plus précis, nous comparons notre approche avec l'approche qui prend directement en compte les tags fournis par le réseau social de l'utilisateur.

5.1. Base de test

Nous avons évalué notre approche sur la base de test *Delicious*. Cette base de données est extraite de (Ivan *et al.*, 2011). La base de données *Delicious* contient le réseau égocentrique de chaque utilisateur, des marques-pages des utilisateurs et des tags des utilisateurs. Les utilisateurs U sont décrits par leur ID (Identifiant) par exemple *UserID=8*. Les ressources R sont décrites par leur ID, titre et l'URL par exemple : *1 IFLA - le site Web officiel des Internationaux Fédération d'Associations de Bibliothèque et Institutions http://www.ifla.org/*. Les tags T sont décrits par leur ID et valeur par exemple : *1 developpement*. La base de test contient :

– 1 867 utilisateurs, 7 668 relations bidirectionnelles et une moyenne de 8,236 relations par utilisateur.

– 69 226 URLs dont 38 581 URLs principales, ex. : www.delicious.com

– 53 388 tags, 437 593 tag *assignements* (tas), sous forme de tuples [user, tag, URL], et une moyenne de 234,383 tas par URL et une moyenne de 6,321 tas par tags

– 104 799 bookmarks, une moyenne de 56,132 URL annotées par utilisateur et une moyenne de 1,514 utilisateur annotant une URL.

9. Union des tags des personnes proches.

5.2. Mesures

Dans cette section, nous présentons les mesures utilisées dans notre évaluation.

- Précision et précision moyenne : Nous calculons la précision des intérêts détectés selon les intérêts produits par notre approche et en utilisant les personnes proches (cf. formule 1). La précision $Precision(u)$ pour chaque utilisateur $u \in U$ est calculée selon le nombre de tags précis ($C_u \subset I_u$), qui existent à la fois dans le profil individuel de l'utilisateur et les profils calculés par notre approche, et le nombre total de tags fournis par notre approche (I_u) :

$$Precision(u) = \frac{|Cu|}{|Iu|} \tag{1}$$

Nous calculons également la précision moyenne pour tous les utilisateurs (cf. formule 2) fournie à partir de la formule de précision $Precision(u)$ (cf. formule 1) pour un utilisateur u, où n est le nombre d'utilisateurs (dans notre cas, n = 1867) :

$$Precision_moyenne = \frac{\sum_{i=1}^{n} Precision(u)}{n} \tag{2}$$

La précision moyenne des utilisateurs est calculée selon le réseau égocentrique ou selon les communautés. Le réseau égocentrique est défini comme l'ensemble des utilisateurs connectés explicitement avec un utilisateur donné (Tchuente *et al*., 2013). La définition des communautés [10] est proposée par (Cazabet *et al*., 2010) et utilisé dans (Tchuente *et al*., 2013). Les communautés sont détectées grâce à un algorithme appelé « iLCD » qui a prouvé son utilité à gérer la dynamique des réseaux à grande échelle. Nous utilisons cet algorithme afin de générer des communautés associées à nos données. La communauté contient un ensemble d'utilisateurs qui pourraient également être présents dans le réseau égocentrique de l'utilisateur.

- Boîtes de Tuckey (*box plot*) : Ces boîtes reflètent la distribution des valeurs de précision dans les résultats (selon quatre quantiles). Elles sont plus représentatives qu'une moyenne simple des précisions. Pour chaque méthode de comparaison, l'extrémité supérieure de la ligne continue représente la valeur maximale des valeurs obtenues, tandis que l'extrémité inférieure représente la valeur minimale. Concernant le rectangle, il récupère toutes les valeurs situées entre le premier et le troisième quartile (Q3). Ce sont les valeurs de 25 % des données qui sont situées en dessous du premier quartile (Q1), et 25 % des données qui sont situées au-dessus du troisième quartile. L'écart inter-quartile correspond donc à 50 % des valeurs situées dans la partie centrale de la distribution. Il est donc utilisé comme indicateur de la dispersion.

10. Les communautés sont générées par un algorithme prédéfini (voir (Cazabet *et al*., 2010)), non détaillé dans cet article.

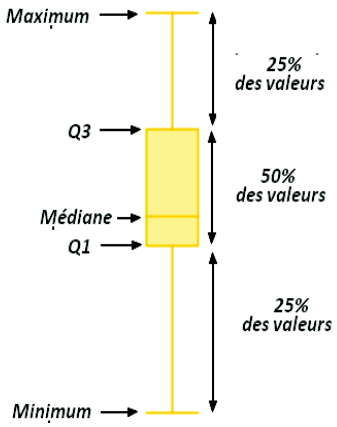


Figure 6. Un exemple de boîtes de Tuckey

5.3. Évaluation selon les personnes proches

Nous testons notre approche sur tous les utilisateurs de la base de données. Ces utilisateurs ont un nombre différent de personnes proches (de 1 à 90 ami(s) explicite(s)). Le nombre de tags, de ressources et de comportement d'annotation est différent pour chaque utilisateur. Ce nombre peut varier d'environ 3 à 800 pour les tags, de 10 à 450 pour les ressources, et de 20 à 500 pour les comportements d'annotation.

5.3.1. Évaluation par rapport à toute la base de test

Notre approche a été testée avec différentes valeurs de k (qui sélectionne les top-k ressources) telles que k = 20, k = 50 et k = 100. Nous calculons la précision moyenne selon le réseau égocentrique et la précision moyenne selon les communautés pour les deux méthodes d'évaluation : « simple comparaison » et « avec synonymes ou mots reliés » (cf. figure 7).

De ces deux tests, nous voyons clairement que la précision qui prend en considération les synonymes ou mots reliés donne de meilleurs résultats que la technique de simple comparaison. Ceci est un résultat attendu puisque les utilisateurs peuvent avoir les mêmes intérêts (tags), mais ils peuvent les décrire différemment en utilisant différents tags.

De plus, nous remarquons que les communautés reflètent d'avantage les intérêts des utilisateurs si nous considérons la technique de simple comparaison. Par contre, le réseau égocentrique reflète d'avantage les intérêts des utilisateurs si nous considérons les synonymes ou mots reliés. Ceci peut être expliqué par l'utilisation des tags complémentaires/similaires pour les personnes qui se connaissent (réseau égocentrique) et donc l'obtention d'une precision meilleure.

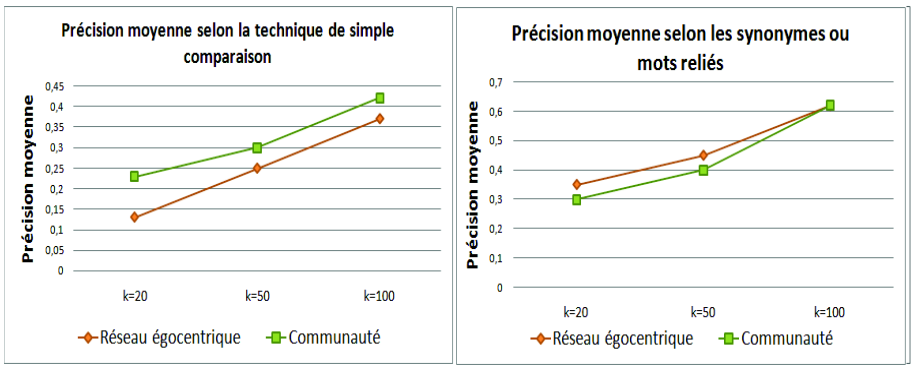


Figure 7. La précision moyenne selon k = 20, k = 50 et k = 100 : selon la technique de simple comparaison (à gauche), selon les synonymes ou mots reliés (à droite)

La valeur de k=100 donne de meilleur résultats. Ceci peut être expliqué par le fait que prendre plus de ressources conduit à une meilleure probabilité de trouver les tags juste au sein de cet ensemble de ressources.

Nous choisissons k = 100 pour le reste de l'évaluation, car comme le montre la figure 7, cette valeur donne de meilleurs résultats. Nous calculons la précision moyenne de tous les utilisateurs dans la base de test selon la technique de « simple comparaison » et « avec synonymes ou mots reliés ». Nous présentons le résultat (pour k=100) à l'aide d'une représentation en boîtes de Tuckey dans la figure 8.

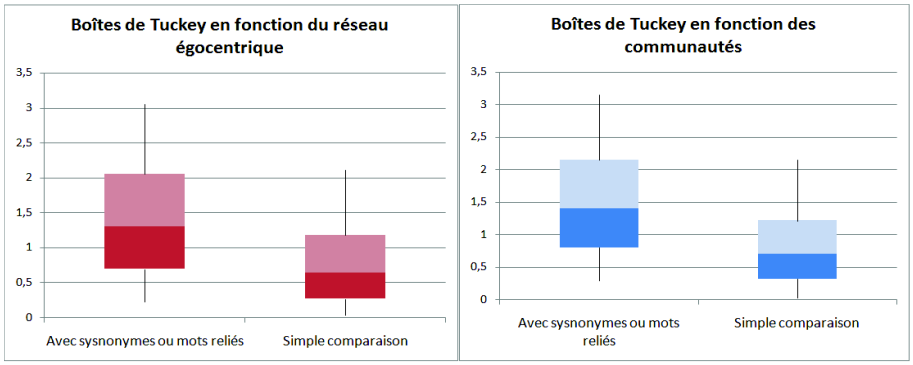


Figure 8. Boîtes de Tuckey de notre approche en fonction du réseau égocentrique (à gauche) et des communautés (à droite)

Nous remarquons que :

– Pour la précision selon les synonymes ou mots reliés, la distribution est presque au milieu à la fois pour le réseau égocentrique et les communautés. Ceci montre que la plupart des utilisateurs ont la même précision moyenne.

– Pour la précision en fonction de la technique de simple comparaison, la distribution est inférieure à la répartition des synonymes ou mots reliés, à la fois pour le réseau égocentrique et les communautés. Ceci montre que la plupart des utilisateurs ont des valeurs de précisions assez basses.

5.3.2. Résultats différenciés selon le type de personnes proches

Afin de mieux comprendre ces résultats, nous détaillons la validation selon le type de personnes proches : le réseau égocentrique et puis les communautés.

Dans notre expérimentation, les personnes proches sont décrites par une relation entre deux utilisateurs (user _i, user _j). Ainsi, dans une première étape, nous analysons le réseau égocentrique utilisateur.

Nous avons choisi de représenter un échantillon de 20 utilisateurs choisis au hasard (nous ne pouvions montrer la précision des 1 876 utilisateurs dans une même figure). Dans cette figure, nous détaillons les différentes valeurs de précision calculées par la technique de simple comparaison et aussi en tenant compte des synonymes ou mots reliés. La figure 9 montre les valeurs de précision pour cet ensemble de 20 utilisateurs. Dans une deuxième étape, nous testons notre approche en prenant en compte

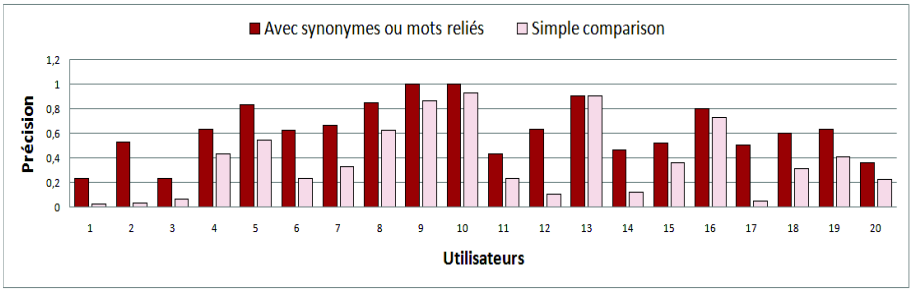


Figure 9. Précision des intérêts détectés pour un ensemble de 20 utilisateurs en fonction de leur réseau égocentrique (k = 100)

les utilisateurs appartenant aux mêmes communautés. La figure 10 montre la précision moyenne des intérêts précis détectés pour le même ensemble de 20 utilisateurs. Dans cette figure, nous détaillons les différentes valeurs de précision calculées par la technique de simple comparaison et aussi en tenant compte des synonymes ou mots reliés.

De cette évaluation (cf. figure 9 et 10), nous voyons clairement que la précision qui prend en considération les synonymes ou mots reliés est généralement meilleure que la technique de simple comparaison. Ceci est un résultat attendu parce que les utilisateurs peuvent avoir les mêmes intérêts, mais ils peuvent les décrire différemment.

Plus généralement, de l'ensemble d'utilisateurs de la base, nous remarquons que la précision (pour les deux méthodes de calcul) varie selon trois cas :

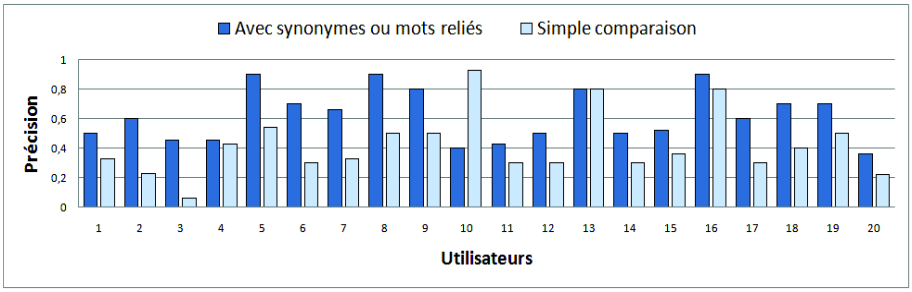


Figure 10. Précision des intérêts précis détectés pour un ensemble de 20 utilisateurs en fonction de leurs communautés (k = 100)

i) la précision est plus élevée pour les utilisateurs actifs (ayant beaucoup de personnes proches et beaucoup de comportements d'annotation). Le nombre de comportement de ces utilisateurs actifs varie de 273 à 1 675, ainsi que le nombre de personnes proches varie de 8 à 90.

ii) la précision est moins élevée pour les utilisateurs moins actifs. Le nombre de comportement de ces utilisateurs moins actifs varie de 1 à 130, ainsi que le nombre de personnes proches varie de 1 à 5.

iii) la précision est égale à zéro (dans les deux cas) lorsque l'écart du nombre des tags fourni par l'utilisateur par rapport à ses personnes proches est important. Par exemple, d'une part, le nombre de comportements d'annotation est faible (ex. : 20) pour un utilisateur donné, et, d'autre part, le nombre de comportements d'annotation de toutes ses personnes proches est important (ex. : 200). Cette différence contribue à réduire les taux de précision.

5.3.3. Élimination des tags non compréhensibles

Nous avons testé si notre approche a éliminé les tags non compréhensibles résultants. Nous notons que les intérêts précis fournis par notre approche sont des mots-clés compréhensibles qui reflètent vraiment le contenu de la ressource comme « technology », « foursquare », « history », etc. Ceci est un avantage, car les tags sont des mots clés générés par les utilisateurs. Notre approche a filtré les tags non compréhensibles (par exemple « gis ») qui ne sont pas compréhensibles par d'autres utilisateurs. Les tags non compréhensibles ont diminué (pour cet ensemble d'utilisateurs) de 35 % à 10 % selon WordNet. Ainsi, l'écart des tags non compréhensibles entre les données d'origine (avant traitement) et les résultats (après traitement) est égal à 71,25 %.

5.4. Évaluation selon l'approche classique basée sur les tags

En utilisant toujours le même ensemble d'utilisateurs, nous avons comparé notre approche avec l'approche classique basée sur les tags. Cette dernière considère les tags

utilisés des utilisateurs comme ses intérêts sans considérer les ressources sur lesquelles ces tags sont posés (Astrain *et al*., 2010) (Li *et al*., 2008).

5.4.1. Évaluation par rapport à toute la base de test

Nous comparons le résultat fourni par notre approche avec le résultat de l'approche classique qui utilise tous les tags des personnes proches (sans tenir compte de leur pertinence par rapport aux ressources associées). Dans ce cas, les intérêts sont calculés à partir de l'union de tous les tags utilisés par les individus proches de l'utilisateur, sans aucun filtrage.

La comparaison est effectuée en fonction du réseau égocentrique de l'utilisateur et aussi en fonction de ses communautés. Nous comparons selon la valeur k = 100 de notre approche (qui fournit les meilleurs résultats). De plus, nous comparons en ne prenant en considération que les synonymes ou mots reliés (puisque cette technique est meilleure que la technique de simple comparaison). Nous calculons la précision pour tous les utilisateurs et nous comparons la précision moyenne fournie par notre approche avec celle fournie par l'approche classique.

A partir de tous les utilisateurs de la base de test, le tableau 1, montre que notre approche permet de surmonter l'approche classique basée sur les tags en termes de précision. Cela est dû à l'examen du contenu des ressources analysées pour la sélection des tags pertinents. Le processus de sélection filtre implicitement les tags non compréhensibles qui peuvent ne pas être compréhensibles pour les autres utilisateurs. Par conséquent, on obtient une précision supérieure à l'approche classique basée sur les tags.

Tableau 1. La précision moyenne de notre approche par rapport à l'approche classique basée tags

	notre approche	**approche basée tags**
Réseau égocentrique	**0,6038**	$0,3459$
Communautés	**0,6125**	$0,3259$

De la même manière que dans la section précédente, nous présentons les boîtes de Tuckey de notre résultat selon les valeurs de précision dans la figure 11.

Cette répartition des valeurs de précision s'explique ainsi :

– Pour la précision selon notre approche, la distribution est presque au milieu à la fois pour le réseau égocentrique et les communautés. Ceci reflète que la plupart des utilisateurs ont la même précision moyenne.

– Pour la précision selon l'approche classique basée sur les tags, la distribution est en dessous de la distribution des synonymes ou des mots reliés, à la fois pour le réseau égocentrique et les communautés. Ceci reflète que la plupart des utilisateurs ont des valeurs de précisions assez basses.

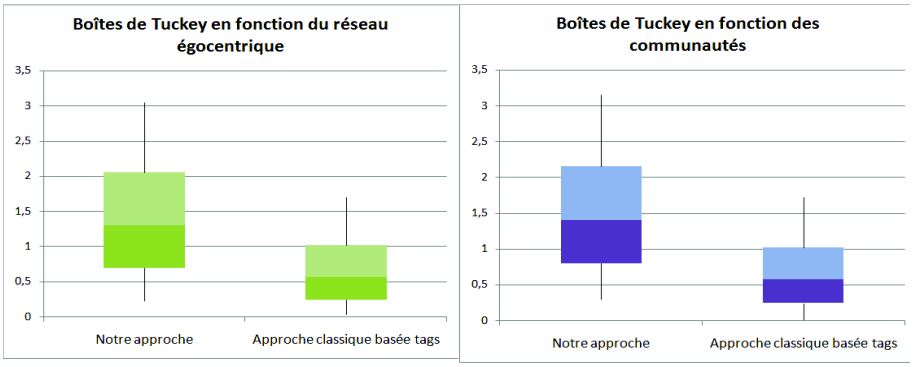


Figure 11. Boîtes de Tuckey de notre approche en fonction du réseau égocentrique (à gauche) et des communautés (à droite)

5.4.2. Résultats différenciés selon le type de personnes proches

Afin de mieux comprendre ces résultats, nous détaillons la comparaison selon le type de personnes proches : le réseau égocentrique puis les communautés.

Nous avons choisi de représenter un échantillon de 20 utilisateurs choisis au hasard (comme nous ne pouvions pas montrer la précision des 1 876 utilisateurs dans une même figure). Pour le réseau égocentrique, la figure 12 compare notre approche avec la précision fournie par l'approche classique basée sur les tags. Pour les communautés, la figure 13 compare notre approche avec la précision fournie par l'approche classique basée sur les tags.

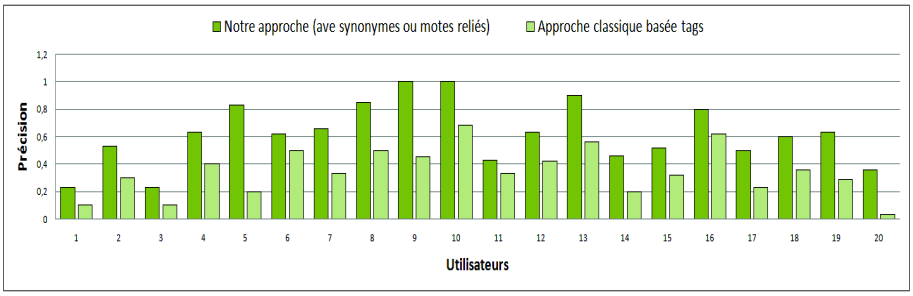


Figure 12. Réseau égocentrique : Comparaison de la précision de notre approche proposée avec la précision fournie par l'approche classique basée sur les tags

De ces comparaisons, nous remarquons que notre approche est généralement plus performante que l'approche basée sur les tags.

De plus, par rapport à tous les utilisateurs de la base, nous constatons que les meilleurs résultats sont liés à des utilisateurs actifs. Le nombre de comportement de ces utilisateurs actifs varie de 273 à 1675, ainsi que le nombre de personnes proches varie de 8 à 90.

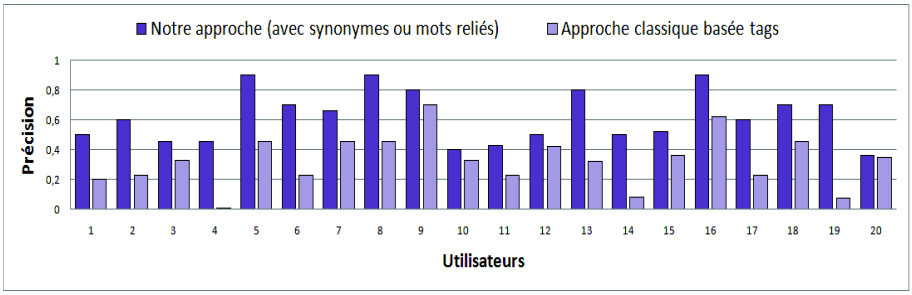


Figure 13. Communautés : Comparaison de la précision de notre approche proposée avec la précision fournie par l'approche classique basée sur les tags

La précision est moins élevée pour les utilisateurs moins actifs. Le nombre de comportement de ces utilisateurs moins actifs varie de 1 à 130, ainsi que le nombre de personnes proches qui varie de 1 à 5.

De plus, pour le réseau égocentrique, la précision est de 88,10 % plus élevée pour notre approche avec les synonymes ou mots reliés que l'approche classique basée sur les tags (pour tous les utilisateurs). Pour les communautés, la précision est de 91,10% plus élevée pour notre approche avec les synonymes ou mots reliés que l'approche classique basée sur les tags (pour tous les utilisateurs).

6. Conclusion

Dans ce papier, nous avons proposé une approche pour détecter les intérêts pertinents des utilisateurs en se basant sur l'environnement social. Le but est de déduire les intérêts des utilisateurs à partir des tags et des relations des utilisateurs. Nous utilisons le contenu des ressources annotées afin de filtrer les tags reflétant vraiment la thématique des ressources et censés représenter les intérêts de l'utilisateur.

L'originalité de notre approche est basée sur la proposition d'une nouvelle technique de détection des intérêts en cherchant à prendre en compte la précision du comportement d'annotation du réseau de relations (réseau égocentrique ou communautés) d'un utilisateur. Cela se fait par l'application d'une technique d'indexation suivie d'une fonction qui score les tags attribués aux ressources. Ce score reflète la pertinence de la ressource par rapport au tag. Ensuite, nous sélectionnons les ressources les plus pertinentes (top-k) pour un tag donné. Si un tag est attribué par un utilisateur à une ressource qui est dans ce top-k, alors ce tag est considéré comme un intérêt précis, c'est-à-dire qu'il décrit le contenu de la ressource à laquelle il a été attribué.

La validation vise à comparer les profils calculés à partir de deux types différents de personnes proches (le réseau égocentrique et les communautés) avec le profil de l'utilisateur. Les résultats ont montré que : i) le réseau égocentrique reflète plus les intérêts des utilisateurs si l'on considère les tags avec les synonymes ou les mots reliés.

ii) les communautés reflètent d'avantage les intérêts des utilisateurs si l'on considère la technique de simple comparaison.

Notre approche fournit un ensemble compréhensible d'intérêts par filtrage des tags réellemet descriptif du contenu des ressources. Par conséquent, notre approche pourrait être utilisée à des fins d'adaptation (par exemple enrichissement du profil utilisateur, recommandation, etc.), car elle offre une solution pour détecter les intérêts pertinents pour les utilisateurs concernés.

Les évaluations réalisées ont montré aussi que la prise en compte des ressources annotées pour détecter les intérêts des utilisateurs concernés (notre approche) est meilleure que de considérer directement les tags attribués par les utilisateurs (approche classique basée sur les tags sans filtrage). En fait, notre approche traite les tags non compréhensibles et donc elle fournit de meilleurs résultats.

Les problèmes qui affectent le processus de détection d'intérêts sont déjà mentionnés dans la section 1. Notre approche a tenté de surmonter ces problèmes de la façon suivante :

– Pour les problèmes de **l'activité dense de l'utilisateur social** et **le manque d'information fournie par l'utilisateur lui-même**, l'approche se focalise sur le comportement d'annotation de son réseau social afin de bénéficier de cette information qui peut refléter les intérêts de l'utilisateur. Nous avons ainsi réduit le spectre d'analyse (analyse du comportement d'annotation des personnes proches seulement). L'information fournie explicitement par l'utilisateur lui-même n'est pas prise en compte.

– Pour le problème de **la variété et la quantité des ressources**, l'approche analyse les personnes proches et principalement le réseau égocentrique et les communautés afin de réduire le spectre d'analyse. De plus, elle se concentre sur l'information textuelle car un rapprochement entre tag et ressource est effectué.

– Pour le problème de **la qualité potentiellement mauvaise des annotations (tags)**, l'approche analyse les tags et leur pertinence par rapport à la ressource associée. Ainsi, les tags ne décrivant pas le contenu sont en grande partie éliminés, et donc les tags non compréhensibles sont réduits.

Les limites de nos travaux qui feront l'objet de travaux futurs sont les suivantes. En ce qui concerne le choix de la valeur de k (qui sélectionne les top-k ressources pour le filtrage des tags), nous envisageons un calcul expérimental par apprentissage afin d'automatiser le choix de cette valeur.

En ce qui concerne la fonction de score (pour le calcul des top-k ressources pour un tag), la limitation principale du modèle utilisé est qu'il ne prend pas en compte les relations entre les mots (par exemple les synonymes). Nous envisageons une prise en compte de la sémantique lors du calcul de score afin de voir l'influence de cette caractéristique sur les valeurs de précisions.

En ce qui concerne les utilisateurs, nous avons montré que notre approche est moins efficace pour les utilisateurs non actifs. Il s'agira de trouver des solutions plus

efficaces que des techniques classiques comme leur attribuer les tags les plus populaires et/ou les plus récents comme étant des intérêts.

Nous comptons également tester notre approche sur d'autres bases sociales, afin de voir son efficacité dans d'autres contextes.

Remerciements

Ce travail a été soutenu financièrement par le programme « PHC Utique » du ministère français des Affaires étrangères et le ministère de l'Enseignement supérieur et la Recherche et le ministère tunisien de l'Enseignement supérieur et la Recherche scientifique sous le numéro de projet CMCU 30540XK.

Bibliographie

Astrain J. J., Cordoba A., Echarte F., Villadangos J. (2010). An algorithm for the improvement of tag-based social interest discovery. In *Semapro '10: Proceedings of the fourth international conference on advances in semantic processing*, p. 49–54. Consulté sur http://www.thinkmind.org/index.php?view=article&articleid=semapro_2010_3_10_50021

Cabanac G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, vol. 87, n° 3, p. 597–620. Consulté sur http://dx.doi.org/10.1007/s11192-011-0358-1

Cai Y., Li Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In *Proceedings of the 19th acm international conference on information and knowledge management*, p. 969–978. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1871437.1871561

Canut C. M., On-at S., Péninou A., Sèdes F. (2015). Enrichissement du profil utilisateur à partir de son réseau social dans un contexte dynamique : application d'une méthode de pondération temporelle. In *Actes du xxxiiie congrès inforsid*, p. 15–30. Consulté sur http://inforsid.fr/Biarritz2015/wp-content/uploads/actes2015/rs-1.pdf

Carmagnola F., Cena F., Cortassa O., Gena C., Torre I. (2007). Towards a tag-based user model: How can user model benefit from tags? In C. Conati, K. McCoy, G. Paliouras (Eds.), *User modeling 2007*, vol. 4511, p. 445-449. Springer Berlin Heidelberg. Consulté sur http://dx.doi.org/10.1007/978-3-540-73078-1_62

Cazabet R., Amblard F., Hanachi C. (2010). Detection of overlapping communities in dynamical social networks. In *2010 IEEE second international conference on social computing (SocialCom)*, p. 309–314.

Godoy D., Amandi A. (2008). Hybrid content and tag-based profiles for recommendation in collaborative tagging systems. In *Latin american web conference, 2008. LA-WEB '08.*, p. 58–65.

Guy I., Zwerdling N., Ronen I., Carmel D., Uziel E. (2010). Social media recommendation based on people and tags. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval*, p. 194–201. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1835449.1835484

Ivan C., Peter B., Tsvi K. (2011). *HetRec '11: Proceedings of the 2Nd international workshop on information heterogeneity and fusion in recommender systems*. New York, NY, USA, ACM.

Kim H.-N., Alkhaldi A., El Saddik A., Jo G.-S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. *Expert Systems with Applications*, vol. 38, n° 7, p. 8488–8496. Consulté sur http://www.sciencedirect.com/science/article/pii/S0957417411000686

Li X., Guo L., Zhao Y. E. (2008). Tag-based social interest discovery. In *Proceedings of the 17th international conference on world wide web*, p. 675–684. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1367497.1367589

Ma Y., Zeng Y., Ren X., Zhong N. (2011). User interests modeling based on multi-source personal information fusion and semantic reasoning. In *Proceedings of the 7th international conference on active media technology*, p. 195–205. Berlin, Heidelberg, Springer-Verlag. Consulté sur http://dl.acm.org/citation.cfm?id=2033896.2033923

Meo P. D., Quattrone G., Ursino D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, vol. 20, n° 1, p. 41–86. Consulté sur http://link.springer.com/article/10.1007/s11257-010-9072-6

Mezghani M., Péninou A., Zayani C. A., Amous I., Sèdes F. (2014). Analyzing tagged resources for social interests detection. In *ICEIS 2014 - proceedings of the 16th international conference on enterprise information systems, volume 1, lisbon, portugal, 27-30 april, 2014*, p. 340–345. Consulté sur http://dx.doi.org/10.5220/0004971303400345

Mezghani M., Péninou A., Zayani C. A., Amous I., Sèdes F. (2015). Détection des intérêts d'un utilisateur par l'exploitation du comportement d'annotation de son réseau égocentrique. In *Actes du xxxiiième congrès inforsid, biarritz, france, may 26-29, 2015*, p. 31–46. Consulté sur http://inforsid.fr/Biarritz2015/wp-content/uploads/actes2015/rs-2.pdf

Milicevic A. K., Nanopoulos A., Ivanovic M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, vol. 33, n° 3, p. 187–209. Consulté sur http://link.springer.com/article/10.1007/s10462-009-9153-2

Musiał K., Kazienko P. (2013). Social networks on the internet. *World Wide Web*, vol. 16, n° 1, p. 31–72. Consulté sur http://link.springer.com/article/10.1007/s11280-011-0155-z

Roth M., Ben-David A., Deutscher D., Flysher G., Horn I., Leichtberg A. *et al*. (2010). Suggesting friends using the implicit social graph. In *Proceedings of the 16th acm sigkdd international conference on knowledge discovery and data mining*, p. 233–242. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1835804.1835836

Song Y., Zhang L., Giles C. L. (2011). Automatic tag recommendation algorithms for social recommender systems. *ACM Trans. Web*, vol. 5, n° 1, p. 4:1–4:31. Consulté sur http://doi.acm.org/10.1145/1921591.1921595

Tchuente D., Canut M.-F., Jessel N., Peninou A., Sèdes F. (2013). A community-based algorithm for deriving users' profiles from egocentrics networks: experiment on facebook and DBLP. *Social Network Analysis and Mining*, vol. 3, n° 3, p. 667–683. Consulté sur http://link.springer.com/article/10.1007/s13278-013-0113-0

Vallet D., Cantador I., Jose J. M. (2010). Personalizing web search with folksonomy-based user and document profiles. In C. Gurrin *et al*. (Eds.), *Advances in information retrieval*, p. 420–

431. Springer Berlin Heidelberg. Consulté sur http://link.springer.com/chapter/10.1007/978-3-642-12275-0_37

Wang J., Clements M., Yang J., Vries A. P. de, Reinders M. J. T. (2010). Personalization of tagging systems. *Inf. Process. Manage.*, vol. 46, n° 1, p. 58–70. Consulté sur http://dx.doi.org/10.1016/j.ipm.2009.06.002

White R. W., Bailey P., Chen L. (2009). Predicting user interests from contextual information. In *Proceedings of the 32Nd international ACM SIGIR conference on research and development in information retrieval*, p. 363–370. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1571941.1572005

Zhang B., Guan Y., Sun H., Liu Q., Kong J. (2010). Survey of user behaviors as implicit feedback. In *2010 international conference on computer, mechatronics, control and electronic engineering (CMCE)*, vol. 6, p. 345–348.

Zheng N., Li Q. (2011). A recommender system based on tag and time information for social tagging systems. *Expert Syst. Appl.*, vol. 38, n° 4, p. 4575–4587. Consulté sur http://dx.doi.org/10.1016/j.eswa.2010.09.131

Zhou T. C., Ma H., Lyu M. R., King I. (2010). Userrec: A user recommendation framework in social tagging systems. In M. Fox, D. Poole (Eds.), *Aaai*. AAAI Press.