Détection des intérêts d'un utilisateur par l'exploitation du comportement d'annotation de son réseau égocentrique

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*
   {mezghanni.manel,andre.peninou,florence.sedes}@irit.fr

RÉSUMÉ. *Les médias sociaux fournissent un environnement d'échange et reposent principalement sur leurs utilisateurs dont le rôle est de créer, d'annoter le contenu des ressources et de construire des relations avec d'autres utilisateurs. Nous nous concentrons sur l'analyse de ces environnements sociaux afin de détecter les intérêts des utilisateurs qui sont des éléments clés pour améliorer l'accès à l'information. L'originalité de notre approche est basée sur la proposition d'une nouvelle technique de détection des intérêts en analysant, d'une part le réseau égocentrique d'un utilisateur, et, d'autre part, la précision du comportement d'annotation des utilisateurs dans le but de sélectionner les tags qui reflètent réellement des intérêts. L'approche proposée a été testée et évaluée sur la base de données sociales Delicious. Une évaluation comparative avec une méthode classique (basée sur les tags) de détection des intérêts montre que notre approche donne de meilleurs résultats.*

ABSTRACT. *Social media provide an environment of information exchange. They principally rely on their users to create, to annotate content and to make on-line relationships. We focus on analysing this social environment to detect user interests which are the key elements for improving adaptation. The originality of our approach is based on the proposal of a new technique of interests detection by analysing the egocentric network of the user and the accuracy of the tagging behaviour 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. A comparative evaluation with the classical tag-based approach 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 a servi dans plusieurs applications comme le e-commerce, le e-learning, etc.
Plusieurs techniques ont été développées afin de mieux se servir de cette connaissance collective. Parmi ces techniques, citons l'adaptation qui permet de fournir à l'utilisateur une information adaptée à ses besoins. L'adaptation est donc un processus fortement lié à l'utilisateur puisqu'on adapte l'information selon ses besoins spécifiques. Un profil utilisateur 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. Nous considérons dans cet article qu'un profil utilisateur est un vecteur de mots-clés (intérêts).
De manière classique, les intérêts d'un utilisateur sont extraits de son propre profil (par exemple de l'attribut intérêt). Toutefois, l'utilisateur généralement ne donne pas toutes les informations relatives à ses intérêts. Donc, le profil explicite de l'utilisateur ne peut jamais être considéré comme entièrement connu par un système. Ainsi, il est difficile de s'appuyer sur la seule analyse du profil pour détecter les intérêts réels (Tchuente *et al.*, 2013).
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 de cet individu peuvent être extraits. Cette hypothèse a été prouvée dans le contexte de dérivation des profils utilisateurs des réseaux sociaux (Tchuente *et al.*, 2013). Les personnes proches peuvent être celles partageant certains comportements communs (par exemple en visitant ou en annotant la même ressource), le réseau égocentrique [1], les utilisateurs appartenant à une même communauté, etc. Donc, nous analysons les personnes proches afin de détecter les intérêts les plus pertinents pour chaque utilisateur. Nous considérons dans ce papier les personnes proches comme celles qui appartiennent au réseau égocentrique d'un utilisateur.
Nous nous concentrons sur le comportement d'annotation des personnes du réseau égocentrique, qui, reflète l'opinion de ces utilisateurs sur une ressource (Astrain *et al.*, 2010). Ce comportement est défini par l'action d'annoter une ressource par un utilisateur. Il est représenté sous la forme de tuples <Utilisateur, Tag, Ressource>. Cette information a prouvé son utilité pour détecter les intérêts des utilisateurs (Meo *et al.*, 2010) (Kim *et al.*, 2011). Un tag est une annotation sociale générée par un utilisateur, qui reflète son intérêt sur une ressource. Malgré l'importance de cette information, elle peut être ambigüe. Nous utilisons aussi le contenu des ressources annotées afin d'éliminer des tags ambigus (opinions, spam, etc.).

L'approche proposée traite principalement les ressources textuelles (semi-structurées, texte, etc.) qui sont présents dans presque tous les principaux médias sociaux tels que :

1. Réseau social d'un utilisateur (égo) réduit aux utilisateurs avec qui il (l'égo) est en relation directe.

Delicious en analysant les URLs annotées, *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*. Nos résultats sont comparés à l'approche utilisant directement les tags fournis par les utilisateurs (approche classique basée sur les tags).

Le reste de cet article est structuré comme suit. Dans la deuxième section, nous présentons notre positionnement par rapport aux éléments d'information sur le comportement d'annotation (utilisateur, tag et ressource). Dans la troisième section, nous présentons et décrivons l'approche proposée qui s'appuie principalement sur un filtrage des tags à partir des ressources et le réseau égocentrique de chaque utilisateur. Dans la quatrième section, nous présentons et commentons les résultats de notre expérimentation sur la base sociale *Delicious*. Dans la dernière section, nous concluons et présentons les perspectives de notre travail.

2. Positionnement

Un système d'adaptation social efficace essaie 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 fait une utilisation sélective de l'information disponible dans l'objectif de construire une liste d'intérêts précise pour chaque utilisateur. Nous détaillons notre analyse par rapport aux éléments du comportement d'annotation ci-dessous :

– **Utilisateur** : Afin de développer notre approche, nous analysons le comportement d'annotation des personnes proches (le réseau égocentrique) de chaque utilisateur. 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), (On-at *et al.*, 2014) et (Zhou *et al.*, 2010).

b) l'absence des informations dans le profil 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 connaitre ses intérêts (Tchuente *et al.*, 2013).

c) l'inefficacité de l'information déduite de son comportement classique. En fait, ce dernier 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) (Mezghani *et al.*, 2012). 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 document 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 détecter des tags significatifs (compréhensibles) plutôt que les tags spécifiques à l'utilisateur.

– **Ressource** : Généralement, les approches traitant cette information utilisent des techniques telles que le *clustering*, le traitement sémantique, etc. (Ma *et al.*, 2011). Cependant, analyser seulement le contenu de la ressource ne permet pas d'avoir des informations sur les intérêts de l'utilisateur (Ma *et al.*, 2011). Le contenu des ressources peut permettre de détecter la nature des tags qui lui sont associés. Cependant, la plupart des recherches ne considèrent pas l'exactitude des tags avec le contenu de la ressource selon (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. Notre approche utilise les tags des personnes proches et les traite en fonction du contenu de leurs ressources respectives. Les tags considérés comme des intérêts pertinents sont ceux qui reflètent le contenu des ressources auxquelles ils ont été associés.

3. Description de l'approche de détection des intérêts

Le processus de détection des intérêts que nous proposons analyse, d'une part le réseau égocentrique d'un utilisateur, et, d'autre part, la précision du comportement d'annotation d'un utilisateur dans le but de sélectionner les tags qui reflètent réellement le contenu des ressources. Le filtrage des tags qui reflètent réellement le contenu des ressources est fait en plusieurs étapes :
i) 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.
ii) Pour chaque tag, attribution d'un score à ces ressources et sélection des top-k ressources.
iii) 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 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 des 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 d'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.

3.1. Préparation des données

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

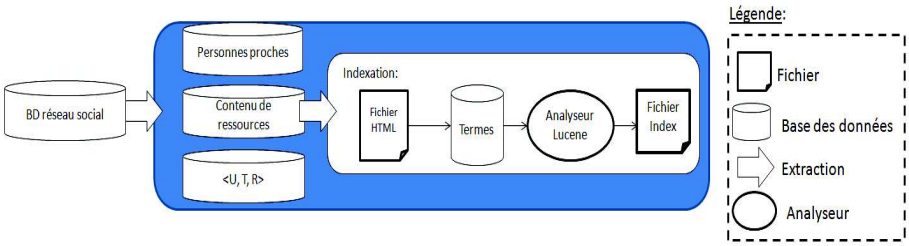

Figure 1. Préparation des données

Nous extrayons les données relatives au :

– Comportement d'annotation $< U, T, R >$, qui est constitués par les tags appliquées aux ressources par les utilisateurs.

– Personnes proches N_u (le réseau égocentrique).

– Contenu des ressources.

Après avoir extrait les données, nous indexons les ressources extraites. L'indexation vise à décrire le contenu d'une ressource par des mots-clés. Les ressources sont indexées, en utilisant *l'API Lucene* [2]. *Lucene* est capable de traiter de grands volumes de ressources 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*, le *URL*, etc.
Les étapes de cette indexation sont les suivantes (cf. figure 1) : *Lucene* indexe les ressources avec un parseur [3] 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 au contenu de la ressource.
L'index est composé de segments, pouvant être considéré comme des sous-index bien qu'ils ne soient pas entièrement indépendants. *Lucene* assigne à chaque ressource de l'index un identifiant unique. 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 ressources contenant ce terme et 4) proximité des termes : la position de chaque terme.

3.2. Processus de détection des intérêts

Le processus de détection des intérêts 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 garder 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.

3.2.1. 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 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.

2. http://lucene.apache.org/core/
3. analyseur syntaxique qui étiquette les mots d'un texte

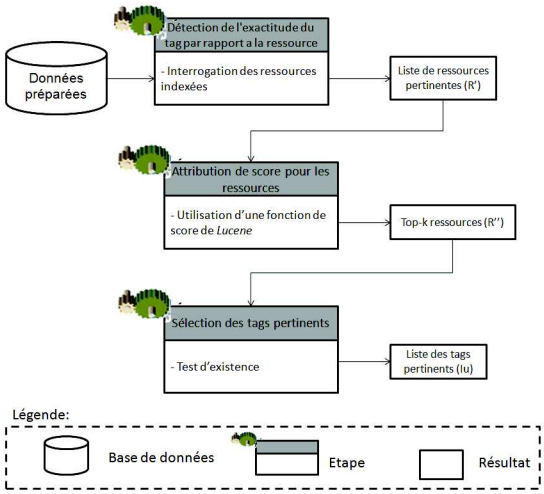

Figure 2. Processus de détection des intérêts pour un utilisateur

3.2.2. 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 de 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 telle que la fonction de similarité soutenue par *Lucene*. Nous choisissons une fonction prédéfinie [4] 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 à cette requête (dans notre cas un tag).
Cette fonction fournit les top-k ressources pertinentes (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 $R'' \subseteq R'$.
La valeur de k est déterminante dans notre cas. Elle est aussi très dépendante de la base de test utilisée. Cette valeur sera choisi dans la partie expérimentation.

3.2.3. 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 fourni 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).

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

Cette étape génère une liste des intérêts pertinents (I_u) sous la forme d'une liste de tags qui décrivent au mieux le contenu de la ressource annotée. Cette liste est issue de l'analyse des personnes proches (le réseau égocentrique) pour chaque utilisateur.

3.3. Processus de validation

Afin de valider notre approche, nous considérons les utilisateurs ayant un profil connu (qui ont déjà eu des activités dans le réseau). Dans une approche classique, nous considérons le profil utilisateur comme étant la liste de tags affectés par l'utilisateur. Donc, nous comparons les tags de l'utilisateur (issus de son profil) avec des tags fournis par notre approche (issus du réseau égocentrique).

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 à travers la figure 3.

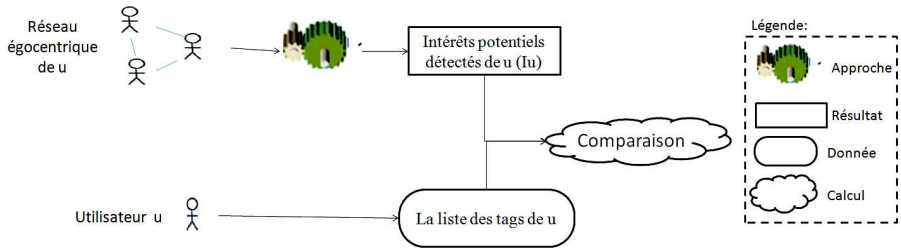

Figure 3. Processus de validation de l'approche de détection des intérêts

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 sont considérés comme l'ensemble de tous les 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 dans le profil utilisateur de u.

La validation de notre proposition, se fait par un test d'existence des intérêts des utilisateurs dans les intérêts potentiels calculés par notre approche. Ce test est effectué à travers deux méthodes :

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

– En prenant en compte les synonymes ou les mots reliés : par exemple si le tag de l'utilisateur= "image" et tag de la personne proche= "photo", alors "image" est considéré comme un tag pertinent. Les synonymes ou mots reliés sont détectés en

interrogeant Wordnet [5]. Nous appellerons cette technique dans le reste de cet article comme "avec synonymes ou mots reliés".

4. Expérimentations sur *Delicious*

Dans cette section, nous détaillons d'abord la base de test utilisée dans la section 4.1. Ensuite, nous présentons les mesures utilisées pour nos calculs dans la section 4.2. Enfin, nous détaillons les évaluations faites afin de tester l'efficacité de notre approche. En effet, notre approche est évaluée selon deux critères :

1. Nous évaluons dans la section 4.3 notre approche selon les deux méthodes de validation : i) comparaison simple ou ii) avec synonymes ou mots reliés. Nous avons testé aussi l'influence de la valeur de k qui sélectionne les top-k ressources pertinents à un tag. Nous conservons les valeurs qui donnent de meilleurs résultats pour faire le reste des évaluations.

2. Nous comparons dans la section 4.4 notre approche avec l'approche qui utilise les informations des tags du réseau égocentrique sans pré-traitement (approche classique basée tags).

Il est à noter que pour l'étape de l'indexation, nous avons pris en compte uniquement le champ *contenu*, vu la richesse de l'information présente.

4.1. 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 [6] 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 :

– 1867 utilisateurs, 7668 relations bidirectionnelles et une moyenne de 8.236 relations par utilisateur.

– 69226 URLs dont 38581 URLs principales, ex. : www.delicious.com

– 53388 tags, 437593 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

– 104799 bookmarks, une moyenne de 56.132 URLs annotées par utilisateur et une moyenne de 1.514 utilisateur annotant une URL.

5. http://wordnet.princeton.edu/

6. Identifiant

4.2. 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 (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 de l'utilisateur et les profils des personnes proches (le réseau égocentrique), et le nombre total des tags trouvés (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) :

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

- Boîtes de Tukey (*box plot*) : Ces boîtes reflètent la distribution des valeurs de précision dans les résultats (selon quatre quartiles). Elles sont plus représentatives qu'une simple moyenne des précisions. Un exemple explicatif de boîtes de Tukey est présenté dans la figure 4. 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). C'est 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 interquartile correspond donc à 50 % des valeurs situées dans la partie centrale de la distribution. Il est donc utilisé comme indicateur de la dispersion.

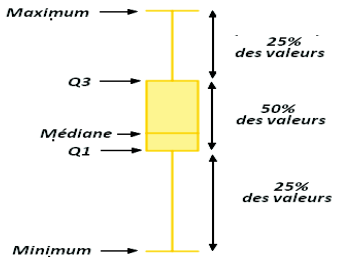

Figure 4. Un exemple de boîtes de Tukey

4.3. Évaluation de notre approche

Nous testons notre approche sur tous les utilisateurs de la base de données. Ces utilisateurs ont un nombre différent de personnes proches (entre 1 et 90 ami(s) explicite(s)). Le nombre de tags, de ressources et de comportement d'annotation (tas) 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.

Notre approche a été testée avec différentes valeurs de k (qui sélectionnent les top-k ressources) tel que k = 20, k = 50 et k = 100. Nous calculons la précision moyenne de notre approche pour les deux méthodes d'évaluation : "simple comparaison" et "avec synonymes ou mots reliés" (cf. figure 5).

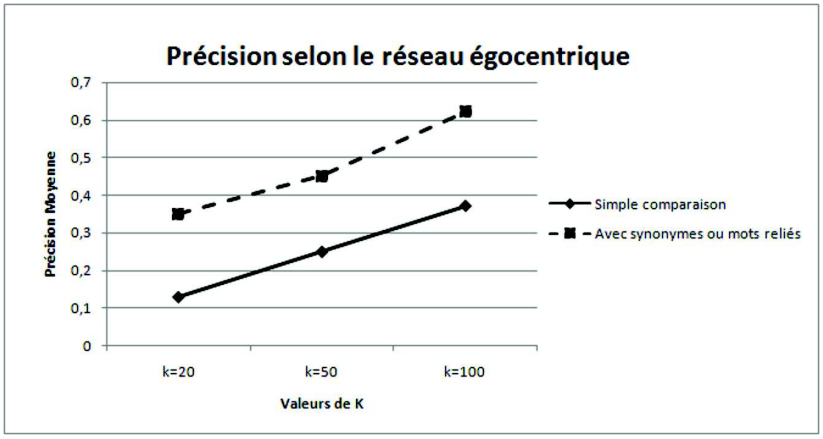

Figure 5. La précision moyenne selon k = 20, k = 50 et k = 100 : (bleu) selon une simple comparaison, (rouge) selon les synonymes ou mots reliés

Nous voyons clairement que la précision qui prend en considération les synonymes ou mots reliés donne de meilleurs résultats que la technique avec simple comparaison (cf. figure 5). Ceci est un résultat attendu puisque les utilisateurs peuvent avoir les mêmes intérêts (tags), mais ils peuvent les décrire différemment en utilisant différents tags.

Nous choisissons k = 100 pour le reste de l'évaluation, car cette valeur donne de meilleurs résultats. Nous présentons le résultat de la précision moyenne selon les deux techniques d'évaluation (pour k=100) à travers une représentation en boîtes de Tukey (*box plot*) dans la figure 6. Nous remarquons que :

– Pour la précision selon les synonymes ou mots reliés, la distribution est presque centrée sur la moyenne. Ceci reflète 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. Ceci reflète que la plupart des utilisateurs ont des valeurs de précisions assez basses.

Afin de comprendre mieux ces résultats, nous avons choisi de représenter un échantillon de 20 utilisateurs choisis au hasard (comme nous ne pouvions pas montrer la précision des 1876 utilisateurs dans une même figure). Dans cette figure, nous détaillons

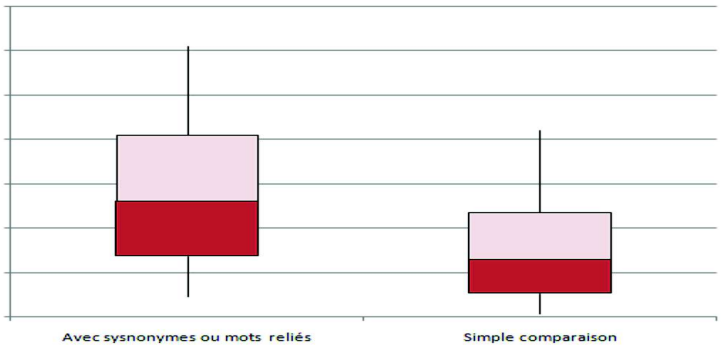

Figure 6. Répartition des précisions selon des boîtes de Tukey de notre approche en fonction du réseau égocentrique

les différentes valeurs de précision calculées par la technique de simple comparaison et aussi en tenant compte des synonymes ou mots reliés. La figure 7 montre les valeurs de précision pour cet ensemble de 20 utilisateurs.

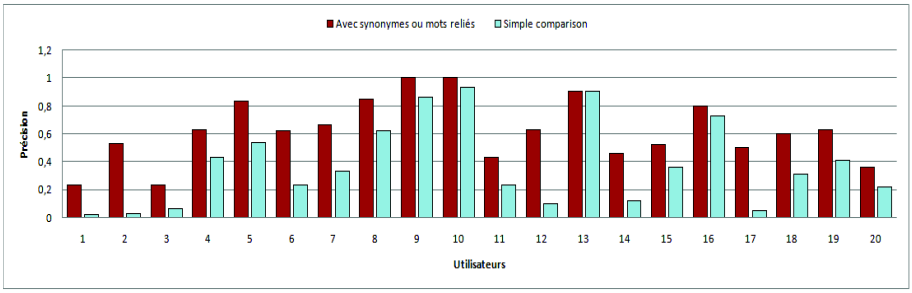

Figure 7. Précision des intérêts détectés pour un ensemble de 20 utilisateurs (k = 100)

De cette évaluation (cf. figure 7), nous voyons clairement que la précision qui prend en considération les synonymes ou mots reliés est généralement meilleure que la technique de simple comparaison.
De tout ensemble d'utilisateurs, nous remarquons que la précision (pour les deux méthodes de calcul) varie selon trois cas : i) la précision est plus élevée pour les utilisateurs actifs (ayant beaucoup de personnes proches et beaucoup de comportements d'annotation). Ces utilisateurs actifs ont une précision de 79% selon la méthode avec synonymes ou mots reliés et de 68% pour la méthode simple comparaison. ii) la précision est moins élevée pour les utilisateurs moins actifs. Ces utilisateurs moins actifs ont une précision de 26% selon la méthode avec synonymes ou mots reliés et de 18% pour la méthode simple comparaison. iii) la précision est égale à zéro 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. D'autre part, le nombre de comportements d'an-

notation de toutes ses personnes proches est important (ex. : 200). Cette différence contribue à réduire les taux de précisions.

De plus, nous avons voulu tester si notre approche a fourni des tags non ambigus. Nous jugeons qu'un tag est compréhensible (resp. ambigu) s'il existe (resp. s'il n'existe pas) dans WordNet. Ainsi 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 ambigus (par exemple "gis") qui ne sont pas compréhensibles par d'autres utilisateurs. L'ambiguïté des tags a diminué (pour cet ensemble d'utilisateurs) de 35 % à 10 % selon WordNet. Ainsi, l'écart d'ambiguïté des tags entre les données d'origine (avant traitement) et les résultats (après traitement) est égal à 71,25 %.

4.4. Évaluation selon l'approche classique basée tags (tag-based)

En utilisant le même ensemble d'utilisateurs que dans la section précédente, nous avons comparé notre approche avec l'approche classique basée sur les tags. Cette dernière considère les tags des utilisateurs comme représentant de ses intérêts (Astrain *et al.*, 2010) (Li *et al.*, 2008).
Nous comparons le résultat fourni par notre approche avec le résultat de l'approche qui utilise tous les tags du réseau égocentrique (sans tenir compte de leur pertinence par rapport aux ressources associées). Nous comparons selon k = 100 de notre approche (fournit les meilleurs résultats). De plus, nous comparons en prenant en considération que les synonymes ou mots reliés (puisqu'elle est meilleure que la simple comparaison). Nous calculons la précision moyenne de tous les utilisateurs dans la base de données et nous la comparons avec la précision moyenne fournie par notre approche. Nous avons obtenu une précision moyenne égale à **0.6038** pour notre approche et une précision moyenne égale à **0.3459** pour l'approche classique basée tags. Ceci, montre que notre approche permet de surmonter l'approche classique basée sur les tags en terme 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 ambigus qui peuvent ne pas être compréhensibles pour les autres utilisateurs.

De même que dans la section précédente, nous présentons les boîtes de Tukey de ces résultats selon les valeurs de précision dans la figure 8.

Cette répartition des valeurs de précision s'explique par :

– Pour la précision de notre approche selon les synonymes ou mots reliés, la distribution est presque au milieu. Ceci reflète que la plupart des utilisateurs ont la même précision moyenne.

– Pour la précision selon l'approche classique basée les tags, la distribution est en dessous de la distribution des synonymes ou des mots reliés. Ceci reflète que la plupart des utilisateurs ont des valeurs de précisions assez basses.

Figure 8. Répartition des précisions selon des boîtes de Tukey de notre approche et de l'approche basée tags en fonction du réseau égocentrique

Afin de mieux comprendre ces résultats, nous avons choisi pour représenter un échantillon de 20 utilisateurs choisis au hasard (comme nous ne pouvions pas montrer la précision des 1876 utilisateurs dans une même figure). En fait, les résultats de notre approche ont fourni des échantillons similaires. La figure 9 compare notre approche proposée avec la précision fournie par l'approche basée sur les tags.

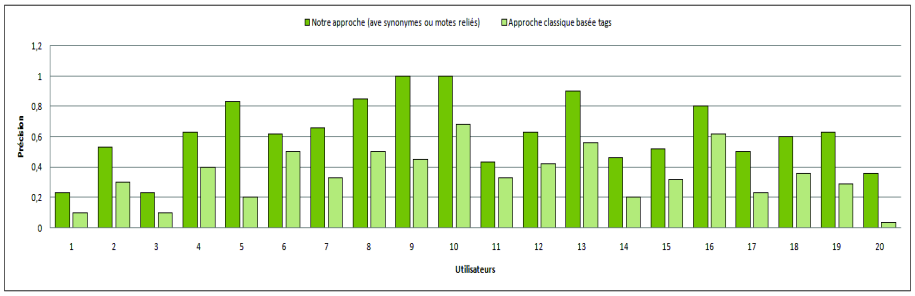

Figure 9. Comparaison de la précision de notre approche proposée avec la précision fournie par l'approche classique basée sur les tags

Nous remarquons que notre approche est généralement plus performante que l'approche basée sur les tags (cf. figure 9). En fait, la précision est 88,10 % plus élevée pour notre approche avec les synonymes ou mots reliés que l'approche basée sur les tags (pour tous les utilisateurs). De plus, nous constatons que les meilleurs résultats sont liés à des utilisateurs actifs.

5. Conclusion

Dans ce papier, nous avons proposé une approche pour détecter les intérêts des utilisateurs précis en se basant sur l'environnement social. Le but est de déduire les intérêts des utilisateurs à partir des tags. Pour cela, 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 égocentrique 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 le tag est attribué par l'utilisateur à une ressource qui est dans ce top-k, alors le tag est considéré comme un intérêt précis, c'est à dire il décrit le contenu des ressources auxquelles il a été attribué.

Les résultats ont montré que notre approche est plus performante lorsqu'elle est complétée par la prise en compte des synonymes ou mots reliés.
De plus, les résultats ont prouvé 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). En fait, notre approche traite implicitement l'ambiguïté des tags et donc elle fournit de meilleurs résultats.
Les expérimentations ont montré que notre approche fournit un ensemble compréhensible d'intérêts. Par conséquent, notre approche pourrait être utilisée à des fins d'adaptation (par exemple la recommandation), car elle offre une solution pour détecter les intérêts des utilisateurs concernés.

Les limites dégagées 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), 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 de ce modèle est qu'il ne prend pas en compte les relations entre les mots (par exemple, les synonymes). Nous pouvons prévoir donc 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 récentes 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

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., Zayani C. A., Amous I., Gargouri F. (2012). A user profile modelling using social annotations: A survey. In *Proceedings of the 21st international conference companion on world wide web*, p. 969–976. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2187980.2188230

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

On-at S., Canut C. M., Péninou A., Sèdes F. (2014). Deriving user's profile from sparse egocentric networks: Using snowball sampling and link prediction. In *Ninth international conference on digital information management, ICDIM 2014, phitsanulok, thailand, september 29 - oct. 1, 2014*, p. 80–85. Consulté sur http://dx.doi.org/10.1109/ICDIM.2014.6991421

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

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.