Étude du profil utilisateur pour la recommandation dans les folksonomies

Mohamed Nader Jelassi, Sadok Benyahia, Mephu Nguifo Engelbert

Étude du profil utilisateur pour la recommandation dans les folksonomies

Mohamed Nader Jelassi[1,2,3], Sadok Ben Yahia[1] et Engelbert Mephu Nguifo[2,3]

[1] Université Tunis El Manar. Faculté des Sciences de Tunis, Tunis, Tunisie.

[2] Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France.

[3] CNRS, UMR 6158, LIMOS, F-63171 Aubière, France.
{nader.jelassi@isima.fr, sadok.benyahia@fst.rnu.tn, engelbert.mephu_nguifo@univ-bpclermont.fr}

Résumé : Dans les *folksonomies*, les utilisateurs partagent des ressources (films, livres, sites web, etc.) en les annotant avec des tags librement choisis. Dans ce papier, nous considérons une nouvelle dimension dans une *folksonomie* qui contient des des informations supplémentaires sur les utilisateurs. Nous définissons un degré de proximité entre deux utilisateurs comme le nombre d'informations de profil en commun entre eux et nous proposons un système personnalisé de recommandations basé sur cette définition. Les expérimentations menées sur un jeu de données du monde réel, MOVIELENS, montrent l'utilité de la nouvelle dimension introduite et quelles informations sont les plus influentes durant le processus de recommandation.

Mots-clés : Folksonomie, Recommandation, Qualité, Précision, Informations supplémentaires, Profil

1 Introduction et Motivations

Une *folksonomie* désigne un système de classification collaborative par les internautes [1] [2] (Mika (2007)). Elle est composée de trois ensembles : un ensemble $\mathcal{U}$ d'utilisateurs, un ensemble $\mathcal{T}$ de tags (ou mots-clés) et un ensemble $\mathcal{R}$ de ressources (films, livres, sites web, photos, etc.) ((Hotho *et al.*, 2006)). Les utilisateurs sont les responsables du partage des ressources et l'affectation de tags à ces derniers (Strohmaier *et al.* (2012)). Cependant, il s'avère que le choix de tags et de ressources partagés par un utilisateur d'une *folksonomie* varie selon plusieurs critères : le genre, l'âge ou encore la profession de celui qui partage l'information. Cela a motivé les chercheurs à proposer des systèmes de recommandation personnalisés afin de répondre aux besoins de chaque utilisateur selon son profil. Ainsi, un système de recommandation personnalisé offre à l'utilisateur des tags et ressources en respectant le profil de ce dernier de telle sorte que les recommandations soient le plus proche de ses besoins (Ricci *et al.* (2011)) ((Nanopoulos *et al.*, 2010)). Pour répondre à cette tâche, nous considérons, dans ce papier, une quatrième dimension dans une *folksonomie*. Cette quatrième dimension peut recouvrir différents aspects : par exemple le profil (genre, âge, profession, ...) comme mentionné ci-dessus, ou le temps si on veut étudier la dynamique temporelle des *folksonomies*. Dans ce papier, nous traitons la quatrième dimension de manière indifférente pour l'aspect méthodologique, mais afin d'avoir des informations disponibles dans un jeu de données du monde réel et d'étudier l'influence de ces informations, nous focaliserons sur l'aspect profil. Par ailleurs, nous définissons un degré de proximité entre deux utilisateurs correspondant au nombre de

1. `http://www.vanderwal.net/folksonomy.html`

2. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`

leurs variables (informations de profil, date de partage, etc.) en commun. Ainsi, grâce au système personnalisé de recommandation, que nous introduisons, qui repose sur cette mesure, nous montrons l'utilité des variables introduites dans la quatrième dimension durant le processus de recommandation. Enfin, nous menons des expérimentations sur un jeu de données du monde réel, *i.e.,* MOVIELENS, pour savoir quelles informations sont les plus influentes pour la recommandation. De plus, nous comparons les précisions de notre système de recommandation, avec et sans considération d'informations supplémentaires, aux approches de la littérature, et nous répondons aux questions suivantes : *(i)* est-ce que la dimension introduite est une information importante aidant à améliorer les recommandations dans les *folksonomies* ? ; *(ii)* à quel point la nouvelle dimension peut-être utile pour fournir des recommandations avec une meilleure précision ? ; *(iii)* et quelles informations supplémentaires sont les plus influentes durant le processus de recommandation ?

Le reste du papier est organisé comme suit : dans la section suivante, nous étudions les principales approches de la littérature. Nous présentons les notions de base dans la Section 3. Ensuite, dans la Section 4, nous introduisons notre système personnalisé de recommandation. Dans la Section 5, nous discutons les résultats de notre étude expérimentale. Enfin, nous concluons notre papier avec des perspectives pour nos travaux futurs dans la Section 6.

2 Travaux connexes

Dans un souci d'améliorer les recommandations dans les *folksonomies*, plusieurs travaux ont été proposés dans la littérature ((De Meo *et al.*, 2010)) ((Basile *et al.*, 2007)) ((Liang *et al.*, 2010)). Dans (Diederich & Iofciu (2006)), les auteurs utilisent la *"personomie"* d'un utilisateur, *i.e.,* les tags qui lui sont relatifs, afin de lui recommander des utilisateurs ayant partagé des tags et ressources similaires. Tout d'abord, ils construisent un profil pour chaque utilisateur. Ensuite, à partir de ce profil, les auteurs sont capables de recommander des utilisateurs (dits *collaborateurs*) en utilisant une mesure de similarité entre utilisateurs. Cette mesure, qui s'appuie uniquement sur les tags utilisés par les utilisateurs, n'offre pas une information complète sur les utilisateurs. Plus récemment, dans (Hu *et al.* (2011)), les auteurs se basent à la fois sur l'historique de tagging (tags et ressources) des utilisateurs et sur leurs contacts sociaux. La limite de cette approche est qu'elle requiert qu'un utilisateur doit posséder des contacts sociaux afin d'avoir des recommandations de tags. Dans (Jäschke *et al.* (2007)), Hotho *et al.* ont proposé des recommandations de tags dans les *folksonomies* basées sur les tags les plus utilisés. Cependant, ces recommandations ne sont absolument pas personnalisées étant donné que les mêmes tags sont proposés à chaque utilisateur. Lipczak a proposé dans (Lipczak (2008)) un système de recommandation de tags en trois étapes. À partir des tags annotés aux ressources, l'auteur ajoute des tags proposés par un lexique basé sur les co-occurrences de tags sur les mêmes ressources. Ensuite, le système filtre les tags déjà utilisés par l'utilisateur. Toutefois, malgré cette étape de filtrage, la recommandation ne paraît pas être personnalisée étant donné qu'elle cherche des tags co-occurrant sur d'autres annotations. L'approche revient ensuite à enlever les tags précédemment annotés par l'utilisateur de ceux qui sont suggérés. Dans (Landia & Anand (2009)), les auteurs ont proposé une nouvelle approche combinant la similarité à la fois entre ressources et entre utilisateurs afin de recommander des tags personnalisés. En effet, deux utilisateurs sont considérés comme similaires s'ils ont assigné les mêmes tags aux mêmes ressources. Toutefois, il est rare de trouver pareille situation dans des *folksonomies* où les tags utilisés par deux

utilisateurs sur les mêmes ressources sont identiques.

Dans notre approche, nous insistons sur le nécessaire recours à des informations supplémentaires et à les combiner à l'historique de tagging afin d'améliorer les recommandations. Toutes ces informations seront représentées par des quadri-concepts. Ainsi, dans ces structures, nous nous focalisons non seulement sur les tags/ressources les plus utilisés, mais également sur ceux qui ont été utilisés en combinaison par des utilisateurs *proches*, obtenant ainsi un résultat plus spécifique. Contrairement aux approches de la littérature qui se limitent à l'information <utilisateur, tag, ressource>, nous étendons ce triplet par l'information contenue dans la quatrième dimension. De plus, les quadri-concepts sont une représentation condensée, sans perte d'information, d'une *folksonomie* dont les données sont souvent altérées par des tags redondants ou par des utilisateurs inactifs.

Dans ce qui suit, nous présentons quelques notions qui seront utilisées tout au long de ce papier.

3 Notions de base

Nous commençons par présenter une extension de la notion de *folksonomie* (Jäschke *et al.* (2008)) par l'ajout d'une quatrième dimension ((Jelassi *et al.*, 2015)).

Définition 1
Une ***v-folksonomie*** *est un ensemble de tuples* $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$ *où* $\mathcal{U}$, $\mathcal{T}$, $\mathcal{R}$ *et* $\mathcal{V}$ *sont des ensembles finis dont les éléments sont appelés* ***utilisateurs***, ***tags***, ***ressources*** *et* ***variables****.* $\mathrm{Y} \subseteq \mathcal{U} \times \mathcal{T} \times \mathcal{R} \times \mathcal{V}$ *représente une relation quadratique où chaque élément* $y \subseteq \mathrm{Y}$ *peut être représenté par un quadruplet :* $y = \{(u, t, r, v) \mid u \in \mathcal{U}, t \in \mathcal{T}, r \in \mathcal{R}, v \in \mathcal{V}\}$ *ce qui veut dire que l'utilisateur* u *a annoté la ressource* r *via le tag* t *à travers la variable* v*. Nous considérons que deux utilisateurs sont* proches *s'ils partagent au moins une même variable en commun.*

La quatrième dimension introduite peut recouvrir différents aspects : par exemple le profil (genre, âge, profession, ...), ou le temps si on veut étudier la dynamique temporelle des *folksonomies*. Ainsi, l'information inclue dans la quatrième dimension est complètement corrélée au triplet (utilisateur, tag, ressource). Par exemple, dans un quadruplet (u, t, r, *date*), l'information *date* est corrélée à l'opération de tagging faite par u avec le tag t sur la ressource r. Dans un autre quadruplet (u, t, r, *profil*), l'information *profil* est corrélée aussi bien à l'utilisateur u qu'au tag t et la ressource r partagés par u. En effet, un utilisateur u peut partager un livre sur les langages de programmation avec le tag *programming* via le profil p_1 (*étudiant* par exemple) tandis qu'il peut partager un papier d'une revue scientifique avec le tag *paper* via le profil p_2 (*chercheur* par exemple). Dans ce papier, nous traitons la quatrième dimension de manière indifférente pour l'aspect méthodologique, mais afin d'avoir des informations disponibles dans un jeu de données du monde réel, nous focaliserons sur l'aspect profil comme quatrième dimension de la *folksonomie*.

Exemple 1
Le Tableau 1 montre un exemple d'une v-folksonomie $\mathcal{F}_v$ *avec* $\mathcal{U} = \{u_1, u_2, u_3, u_4\}$, $\mathcal{T} = \{t_1, t_2, t_3, t_4\}$, $\mathcal{R} = \{r_1, r_2, r_3\}$ *et* $\mathcal{V} = \{v_1, v_2\}$*. Chaque croix d'une relation quadratique, indique une opération de tagging faite par un utilisateur de l'ensemble* $\mathcal{U}$ *avec une variable de* $\mathcal{V}$*, utilisant*

un tag de $\mathcal{T}$ sur une ressource de $\mathcal{R}$. Par exemple, l'utilisateur u_1, qui a entre 25 et 35 ans et qui est étudiant, a taggué les articles r_1, r_2 et r_3 avec les tags thesis, web et to_recommend.

$\mathcal{F}_v$	$\mathcal{R}$	r_1				r_2				r_3			
$\mathcal{V}$	$\mathcal{U}/\mathcal{T}$	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
	u_1		×	×	×		×	×	×		×	×	×
v_1	u_2		×	×	×	×	×	×	×	×	×	×	×
	u_3		×	×	×	×	×	×	×	×	×	×	×
	u_4		×	×		×			×	×			×
	u_1		×	×	×		×	×	×		×	×	×
v_2	u_2		×	×	×	×			×	×	×	×	×
	u_3												
	u_4												

TABLE 1 – Un exemple d'une *v-folksonomie* avec les valeurs suivantes : *science*(t_1), *thesis*(t_2), *web*(t_3), *to_recommend*(t_4), *25-35 ans*(v_1), *étudiant*(v_2) et r_1, r_2 et r_3 trois articles scientifiques.

Nous définissons maintenant un concept quadratique (Jelassi *et al.* (2015)).

Définition 2
*Un concept quadratique (ou quadri-concept) d'une v-folksonomie $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$ est un quadruplet (U, T, R, V) avec $U \subseteq \mathcal{U}$, $T \subseteq \mathcal{T}$, $R \subseteq \mathcal{R}$ et $V \subseteq \mathcal{V}$ avec $U \times T \times R \times V \subseteq \mathrm{Y}$ tel que le quadruplet (U, T, R, V) est maximal, i.e., aucun de ces ensembles ne peut être augmenté sans diminuer un des trois autres ensembles. Pour un quadri-concept $QC = (U, T, R, V)$, U, R, T et V sont, respectivement, appelés **Extent**, **Intent**,* **Modus** *et **Variable**.*

Remarque 1
Afin de permettre l'extraction de l'ensemble de quadri-concepts fréquents à partir d'une v-folksonomie *donnée, nous pouvons utiliser l'un des deux algorithmes de la littérature dédiés à cette tâche :* QUADRICONS *(Jelassi* et al. *(2013)) ou* DATAPEELER *(Cerf* et al. *(2009)). Les deux algorithmes prennent en entrée une* v-folksonomie *ainsi que quatre seuils minimaux de support (un pour chaque dimension) et donnent en sortie l'ensemble de **quadri-concepts** vérifiant ces seuils. Un quadri-concept **fréquent** est un quadri-concept dont chaque ensemble (utilisateur, tag, ressource et variable) a une cardinalité supérieure ou égale à son seuil de support correspondant.*

Définition 3
*(*DEGRÉ DE PROXIMITÉ*) Considérons une v-folksonomie $\mathcal{F}_v = (\mathcal{U}, \mathcal{T}, \mathcal{R}, \mathcal{V}, \mathrm{Y})$, nous définissons le **degré de proximité** entre deux utilisateurs de $\mathcal{U}$ comme le nombre de leurs variables de $\mathcal{V}$ en commun.*

Exemple 2
Considérons la v-folksonomie, *représentée par le Tableau 1, le quadri-concept ($\{u_1, u_2, u_3\}$, $\{t_2, t_3, t_4\}$, $\{r_1, r_2, r_3\}$, v_1) montre que les utilisateurs u_1, u_2 et u_3 ont un degré de proximité égal à* 1, i.e., *ils partagent la variable v_1 en commun. Par contre, le quadri-concept ($\{u_1, u_2\}$, $\{t_2$,*

t_3, t_4}, {r_1, r_3}, {v_1, v_2}) *montre que les utilisateurs* u_1 *et* u_2 *ont deux variables en commun,* i.e., v_1 *et* v_2. *Ainsi, ils ont un degré de proximité égal à* 2.

4 Recommender : un nouvel algorithme pour des recommandations personnalisées

Dans cette section, nous proposons notre nouveau système personnalisé de recommandation RECOMMENDER. Le pseudo code de RECOMMENDER est présenté par l'Algorithme 1. RECOMMENDER prend en entrée $\mathcal{QC}$ (un ensemble de quadri-concepts fréquents) ainsi qu'un utilisateur cible u avec son ensemble de variables V, un degré de proximité d et (optionellement) une ressource r (à annoter). L'ensemble $\mathcal{QC}$ est extrait dans une étape de pré-traitement par l'un des algorithmes de la littérature dédiés à cette tâche. Ensuite, RECOMMENDER donne en sortie trois ensembles : un ensemble d'utilisateurs proposés, un ensemble de tags suggérés et un ensemble de ressources recommandées en prenant en considération le degré de proximité d entre utilisateurs. Ce degré est une métrique définie par l'utilisateur et qui est égale, au minimum, à 0 et, au maximum, aux nombre de variables disponibles dans le jeu de données considéré.

RECOMMENDER opère comme suit : il commence par initialiser tous les ensembles de sortie aux ensembles nuls (Ligne 2). Ensuite, il extrait les tags et ressources déjà partagés par l'utilisateur u (Lignes 3-4) afin d'éviter de lui recommander des tags et des ressources qu'il a déjà partagé. Par suite, selon le degré de proximité, RECOMMENDER opère comme suit : (*i*) si la valeur de d est égale à 0 (Lignes 5-12), *i.e.,* la recommandation est indépendante de l'ensemble de variables V, alors, nous recommandons à u un ensemble d'utilisateurs $\mathcal{PU}$ qui ont partagé les mêmes tags et ressources que lui (Ligne 9). De plus, nous recommandons à u un ensemble de ressources partagés par les utilisateurs de l'ensemble $\mathcal{PU}$ (Ligne 10). Enfin, nous recommandons également à u un ensemble de tags utilisés par ces mêmes utilisateurs sur la ressource ra que u souhaite partager (Lignes 11 et 12) ; (*ii*) si la valeur du degré de proximité est égal, au moins, à 1, *i.e.,* l'utilisateur u doit partager, au moins, d variables en commun avec les autres utilisateurs de la *v-folksonomie* (Ligne 15). En parcourant les quadri-concepts de l'ensemble $\mathcal{QC}$, si u appartient déjà à un quadri-concept qc, alors qc est élagué (Ligne 16) afin de filtrer les tags et ressources déjà partagés par u. Cette stratégie est inspirée par celle de (Lipczak (2008)). Ensuite, selon la tâche à accomplir, RECOMMENDER fonctionne comme suit : pour la tâche de *Proposition d'utilisateurs* (Ligne 7), c'est la partie *utilisateurs* du quadri-concept qc qui est ajoutée à l'ensemble $\mathcal{PU}$ des utilisateurs proposés. Cette tâche aide à connecter les utilisateurs qui ont des intérêts communs et aide également à promouvoir le partage de ressources. Pour la tâche de *Suggestion de tags* (Lignes 19 et 20), le but est de suggérer des tags personnalisés à un utilisateur qui souhaite ajouter une ressource à la *folksonomie*. Cette tâche a plusieurs avantages : elle rappelle à l'utilisateur ce dont une ressource s'agit, accroît l'annotation des ressources et permet de consolider le vocabulaire des utilisateurs (Ricci *et al.* (2011)). Pour cette tâche, nous ajoutons donc les tags affectés à la ressource ra par les utilisateurs qui ont d variables en commun avec u à l'ensemble $\mathcal{ST}$. Quant à la tâche de *Recommandation de ressources* (Ligne 18), le but est de proposer une liste personnalisée de ressources conforme aux intérêts de l'utilisateur u ; ces ressources sont ajoutées à l'ensemble $\mathcal{RR}$.

Algorithme 1 : RECOMMENDER

Données :

1. $\mathcal{QC}$: un ensemble de quadri-concepts fréquents
2. u : un utilisateur cible avec son ensemble de variables V
3. d : un degré de proximité
4. ra : une resource à annoter par u

Résultats :

1. $\mathcal{PU}$: un ensemble d'utilisateurs proposés
2. $\mathcal{ST}$: un ensemble de tags suggérés
3. $\mathcal{RR}$: un ensemble de ressources recommandées

```
1  début
2      PU=ST=RR=∅
3      u.Tags={t ∈ T / ∃ r ∈ R ∃ u ∈ U ∃ v ∈ V, (u,t,r,v) est un quadri-concept} ;
4      u.Ressources={r ∈ R / ∃ t ∈ T ∃ r ∈ R ∃ v ∈ V, (u,t,r,v) est un quadri-concept} ;
5      si d=0 alors
6          pour chaque quadri-concept qc ∈ QC faire
7              si u ∈ qc.Extent alors
8                  pour chaque utilisateur u' de qc.Extent faire
9                      PU = PU ∪ u' /*Proposition d'utilisateurs*/
10                     RR = RR ∪ u'.Resources \ u.Resources ; /*Recommandation de
                       ressources*/
11                     si ra ∈ qc.Intent alors
12                         ST = ST ∪ u'.Tags \ u.Tags ; /*Suggestion de tags*/
13     sinon si d > 0 alors
14         pour chaque quadri-concept qc ∈ QC faire
15             si | V ∩ qc.Variable | ≥ d alors
16                 si u ∉ qc.Extent alors
17                     PU = PU ∪ qc.extent /*Proposition d'utilisateurs*/
18                     RR = RR ∪ qc.intent \ u.Resources ; /*Recommandation de
                       ressources*/
19                     si ra ∈ qc.Intent alors
20                         ST = ST ∪ qc.modus \ u.Tags ; /*Suggestion de tags*/
21     retourner (PU,ST,RR) ;
22 fin
```

5 Résultats expérimentaux et Discussion

Dans cette section, nous évaluons notre approche sur un jeu de données du monde réel, *i.e.,* MOVIELENS en calculant la précision de nos recommandations pour différentes valeurs de degré de proximité afin de mettre en valeur l'utilité d'avoir des informations supplémentaires sur les utilisateurs durant le processus de recommandation ((Baeza-Yates & Ribeiro-Neto, 1999)) ((Herlocker *et al.*, 2004)). De plus, nous comparons les différentes précisions obtenues avec notre approche avec les travaux pionniers qui ont un objectif commun avec la nôtre, *i.e.,* ceux de Bellogin *et al.* (Bellogín *et al.* (2013)) et Qumsiyeh *et al.* (Qumsiyeh & Ng (2012)). Ces approches n'utilisent pas de quatrième dimension mais font appel au profil des utilisateurs comme information complémentaire pour la tâche de recommandation.

Le jeu de données filmographique MOVIELENS (*http ://movielens.umn.edu/*) est un système de recommandation et un site web communautaire qui permet aux utilisateurs de partager des films en les annotant par des tags. Le jeu de données, utilisé pour nos expérimentations, est téléchargeable gratuitement (*http ://www.grouplens.org/node/73*) et contient 95580 tags appliqués à 10681 films par 71567 utilisateurs (par exemple, $<$*Alex*, *X-Files*, *sciencefiction*$>$). Le choix du jeu de données MOVIELENS est expliqué par le fait qu'en plus d'être très utilisé dans le domaine de recommandation, ce jeu de données offre des informations supplémentaires sur les utilisateurs : l'âge, la profession ou le genre.

Utilisateur	Tag	Ressource	Profil
Mulder	*action*	*X-Files*	*student*
Mulder	*sciencefiction*	*X-Files*	25 *years old*
Scully	*adventure*	*Jurassic Park*	*professor*
Scully	*bestmovie*	*Jurassic Park*	*female*
Skinner	*thriller*	*Carrie*	*Canada*
⋮	⋮	⋮	⋮

TABLE 2 – Un instantané du jeu de données MOVIELENS.

Afin d'étudier l'influence des informations supplémentaires sur les utilisateurs durant la recommandation, nous avons choisi, dans ce qui suit, le profil des utilisateurs pour modéliser la variable v dans la *v-folksonomie*. Ainsi, nous considérons désormais le degré de proximité entre deux utilisateurs comme le nombre d'informations de profil qu'ils ont en commun (par exemple, le même âge et la même profession, si le degré de proximité est égal à 2). À cet effet, les informations supplémentaires sur les utilisateurs qui sont disponibles dans MOVIELENS sont le **genre** de l'utilisateur (masculin ou féminin), sa **profession** (au nombre de 21, qui peut être éducateur, écrivain, étudiant, scientifique, etc.) ou encore l'**âge** des utilisateurs qui est divisé en cinq tranches : (*i*) $7-18$ ans ; (*ii*) $19-24$ ans ; (*iii*) $25-35$ ans ; (*iv*) $36-45$ ans et (*v*) $46-73$ ans.

Base d'apprentissage/Base de Test

Pour nos expérimentations, nous avons utilisé le protocole de validation "5-validation croisée" ((Weiss & Kulikowski, 1991)) afin d'évaluer la pertinence de notre approche. Le jeu de

données MOVIELENS a été partitionné en deux échantillons : un échantillon aléatoire contenant 80% des utilisateurs a été utilisé comme **base d'apprentissage** et un échantillon aléatoire contenant les 20% d'utilisateurs restants, a été utilisé pour la validation de nos tests (*i.e.,* **base de test**). Pour chaque utilisateur du deuxième échantillon (*i.e.,* utilisateur test), 20% aléatoires de ses tags et ressources sont considérées comme ensemble de test/réponse et 80% comme son ensemble d'apprentissage. Nous avons répété cette expérience cinq fois en changeant à chaque fois les 20% représentant la base de test afin de couvrir les 100% de tout l'ensemble. Pour chaque utilisateur test, notre algorithme de recommandation génére une liste d'éléments (utilisateurs, tags ou ressources) en se basant sur son ensemble d'apprentissage. Si un élément de la liste de recommandation se trouve également dans l'ensemble de test de cet utilisateur, alors l'élément est considéré comme **pertinent**. Pour nos expérimentations, nous avons également fait varier le nombre de recommandations fournies à l'utilisateur : il s'agit des top-k recommandations. Grâce à ça, l'utilisateur peut spécifier les k recommandations les plus pertinentes que le système doit lui retourner. Les k premières réponses sont ceux qui ont les scores les plus élevés (*cf.*, Équation 1).

Score de ranking

Dans le but d'améliorer la précision et le rappel des recommandations proposées dans la littérature, nous proposons un nouveau score de ranking afin de classer les différentes recommandations. Pour un jeu de données donné, les top-k recommandations consistent en une liste d'items classés par valeur de score décroissante. Dans ce qui suit, la fonction de score est définie pour la recommandation de ressource mais peut très bien être définie pour la recommandation de tags ou d'utilisateurs en changeant les variables de l'équation. Ainsi, pour générer une recommandation de ressource pour un utilisateur donné, nous calculons le ranking comme décrit ci-dessus, et nous restreignons les résultats aux top-k premiers résultats (avec les scores les plus élevés). La mesure de score (notée rec_score) correspondant à un ensemble d'informations de profil V est défini comme suit :

$$\textit{rec\_score}(r_i, V) = \frac{|u_i|}{|UU|} / \exists t_i \exists r_i \exists v_i \in V, (u_i, t_i, r_i, v_i) \in \mathcal{F}_v \tag{1}$$

Donc, le score rec_score d'une ressource r_i correspondant à un profil v est le nombre d'utilisateurs uniques, ayant le même profil v (ou au moins une information de profil $v_i \in v$), qui ont partagé cette ressource, divisé par le nombre total d'utilisateurs uniques dans l'ensemble des quadri-concepts fréquents (noté UU). Par exemple, si une ressource r_1 a été partagée par 7 différents utilisateurs (au même profil) parmi une liste de 67 utilisateurs uniques, son score sera égal à 0.104 alors qu'une autre ressource r_2 partagée par 16 différents utilisateurs (au même profil) parmi la même liste aura un score égal à 0.238.

Évaluation des recommandations

Le Tableau 3 montre les valeurs de précision des recommandations obtenues par notre système de recommandation pour différents degrés de proximité et différentes valeurs de k [3] allant

3. le nombre de recommandations retournées à l'utilisateur.

Information de profil / k	6	7	8	9	10	**Précision Moyenne**	**Variance**	**Écart Type**
Degré de proximité = 0								
Aucun	0.56	0.54	0.51	0.48	0.48	0.514	0.000922	0.030364
Degré de proximité = 1								
Âge	0.60	0.57	0.54	0.51	0.50	0.544	0.001447	0.038042
Localisation	0.72	0.73	0.75	0.74	0.71	0.730	0.000200	0.014142
Profession	0.55	0.50	0.51	0.50	0.50	0.512	0.000260	0.016149
Degré de proximité = 2								
Âge + Localisation	0.52	0.52	0.52	0.51	0.51	0.516	0.000019	0.004358
Profession + Localisation	0.53	0.51	0.50	0.44	0.42	0.480	0.001800	0.042426
Âge + Profession	0.63	0.64	0.63	0.64	0.67	0.642	0.000241	0.015543
Degré de proximité = 3								
Âge + Profession + Localisation	0.50	0.42	0.37	0.33	0.30	0.384	0.004938	0.070270
Approches de la littérature								
Bellogin *et al.*	0.40	0.37	0.35	0.33	0.32	0.354	0.000824	0.028705
Qumsiyeh *et al.*	0.27	0.27	0.25	0.24	0.23	0.252	0.000256	0.016000

TABLE 3 – Valeurs de précision des recommandations pour différents degrés de proximité pour le jeu de données MOVIELENS (*cf.,* Figure 1).

de 6 à 10 sur le jeu de données MOVIELENS. Tout d'abord, les résultats démontrent l'utilité de la quatrième dimension, *i.e.,* le profil, durant le processus de recommandation. En effet, les autres meilleurs scores de précision sont atteints lorsque notre système de recommandation prend le profil des utilisateurs comme information supplémentaire. Ainsi, le recours aux informations supplémentaires sur les utilisateurs permet de personnaliser les recommandations et de générer des recommandations plus ciblées. De plus, pour toutes les valeurs de k, notre système de recommandation obtient une meilleure précision que celles de Bellogin *et al.* et Qumsiyeh *et al.* La différence est encore plus grande lorsque nous prenons en compte des informations supplémentaires sur les utilisateurs. Ainsi, RECOMMENDER améliore la précision des approches de Bellogin *et al.* et Qumsiyeh *et al.* de, respectivement, 48.49% et 140.48%. L'information de profil la plus influente est la **localisation** des utilisateurs. En effet, plus les utilisateurs sont proches géographiquement, plus ils ont tendance à partager les mêmes ressources et à avoir le même comportement social selon les traditions culturelles de leurs pays respectifs. Par exemple, les utilisateurs indiens partagent les films de Bollywood tandis que les utilisateurs japonais ont tendance à partager en masse les mangas. Ensuite, la seconde information de profil la plus influente pour la recommandation est l'**âge** des utilisateurs. En effet, les utilisateurs appartenant à la même catégorie d'âge convergent vers un vocabulaire commun (les jeunes utilisateurs contre les anciens utilisateurs) et partagent le même type de ressources, par exemple, les jeunes utilisateurs préfèrent les films d'actions, les plus jeunes partagent les mangas alors que les plus anciens ont tendance à partager les films classiques. Par ailleurs, lorsque nous associons deux informations de profil, l'âge apparaît comme étant l'information de profil la plus importante,

notamment lorsqu'elle est associée à la profession ou la localisation. En effet, la précision de notre système de recommandation augmente sensiblement lorsque nous prenons en compte à la fois l'âge et la profession comme informations supplémentaires étant donné que les utilisateurs d'une même catégorie d'âge et exerçant le même métier ont un profil assez proche, par exemple, des étudiants de [19-24] ans ou encore des techniciens de [25-35] ans.

Toutefois, la précision de nos recommandations atteint ses moins bons résultats lorsque nous combinons toutes les informations de profil. Si la localisation ou la combinaison âge-profession produit de bons résultats en termes de précision, les combiner réduit considérablement la qualité des recommandation. En effet, le nombre de recommandations décroît étant donné qu'il est rare de trouver des utilisateurs ayant à la fois la même profession, le même âge, la même localisation et partageant les mêmes ressources. Comme il es rare de retrouver des utilisateurs ayant ce même genre de profil, les ressources recommandées ont moins de chance d'être pertinentes. Par ailleurs, si l'âge ou la localisation donnent de bons résultats en termes de précision, cela n'est pas le cas pour la profession qui n'est pas une information influente pour nos recommandations. Les utilisateurs exerçant le même métier ne partagent pas forcément les mêmes intérêts. Enfin, lorsque notre système de recommandation ne prend aucune information supplémentaire sur les utilisateurs, la précision décroît rapidement puisque la liste de recommandation est aléatoire, c'est-à-dire, pas personnalisée. Ainsi, dans le cas des *v-folksonomies*, plus le nombre de ressources recommandées augmente, moins elles sont pertinentes. En effet, les ressources les plus partagés par le passé ne sont pas nécessairement partagés dans le futur, donc, le score de précision n'est pas élevé lorsqu'aucune information supplémentaire sur les utilisateurs n'est prise en compte. Nous concluons que prendre en compte des informations supplémentaires sur les utilisateurs permet d'augmenter la précision des recommandations, et pour avoir les meilleurs résultats, il est préférable de s'arrêter à une ou deux informations de profil. Ainsi, si nous prenons une seule information de profil, la localisation et l'âge sont les informations de profil qui donnent les meilleurs scores. Par contre, si nous combinons deux différentes variables, il est conseillé de combiner l'âge avec une autre information de profil.

6 Conclusion et Perspectives

Dans ce papier, nous avons considéré une nouvelle dimension dans une *folksonomie* contenant des informations supplémentaires sur les utilisateurs. Ensuite, nous avons proposé notre système personnalisé de recommandation qui repose sur une mesure de proximité entre utilisateurs afin d'améliorer la qualité des recommandations. Les expérimentations ont montré l'utilité d'avoir des informations supplémentaires durant le processus de recommandation. Parmi nos perspectives de recherche, nous cherchons à étendre les informations supplémentaires aux reviews, commentaires et l'historique de recherche des utilisateurs afin d'avoir un suivi dynamique des utilisateurs et améliorer encore plus les recommandations.

Remerciements.

Ce travail est partiellement financé par le projet franco-tunisien PHC Utique 11G141. Nous remercions les relecteurs anonymes pour leurs remarques constructives.

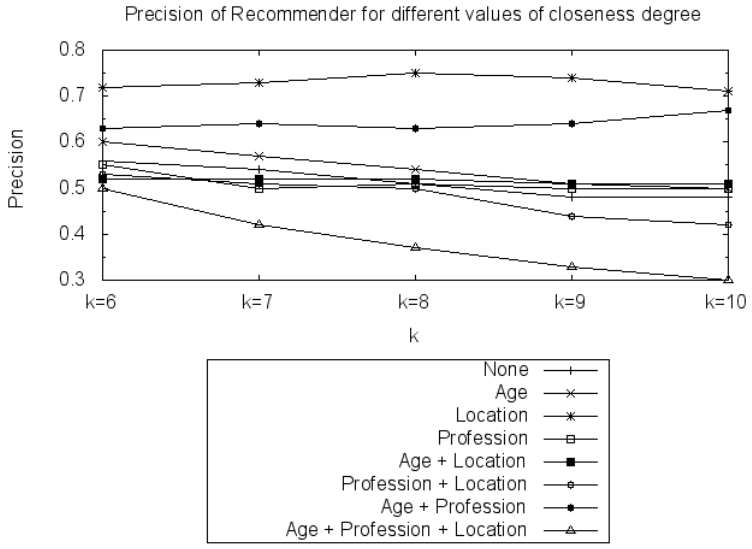

FIGURE 1 – Valeurs de précision des recommandations pour différents degrés de proximité pour le jeu de données MOVIELENS.

Références

BAEZA-YATES R. A. & RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc.

BASILE P., GENDARMI D., LANUBILE F. & SEMERARO G. (2007). Recommending smart tags in a social bookmarking system. In *Bridging the Gap between Semantic Web and Web 2.0*, p. 22–29.

BELLOGÍN A., CANTADOR I. & CASTELLS P. (2013). A comparative study of heterogeneous item recommendations in social systems. *Inf. Sci.*, **221**, 142–169.

CERF L., BESSON J., ROBARDET C. & BOULICAUT J.-F. (2009). Closed patterns meet n-ary relations. *ACM TKDD*, **3**, 3 :1–3 :36.

DE MEO P., 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*, **20**(1), 41–86.

DIEDERICH J. & IOFCIU T. (2006). Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on TEL-CoPs, Crete, Greece*, p. 288–297.

HERLOCKER J. L., KONSTAN J. A., TERVEEN L. G. & RIEDL J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, p. 5–53.

HOTHO A., JÄSCHKE R., SCHMITZ C. & STUMME G. (2006). Information retrieval in folksonomies : Search and ranking. In *Proc. of ESWC, Budva, Montenegro*, volume 4011 of *LNCS*, p. 411–426 : Springer, Heidelberg.

HU J., WANG B. & TAO Z. (2011). Personalized tag recommendation using social contacts. In *Proc. of Workshop SRS'11, in conjunction with CSCW*.

JÄSCHKE R., HOTHO A., SCHMITZ C., GANTER B. & STUMME G. (2008). Discovering shared conceptualizations in folksonomies. *Web Semantics.*, **6**, 38–53.

JÄSCHKE R., MARINHO L., A. HOTHO A., LARS S.-T. & STUMME G. (2007). Tag recommendations in folksonomies. In *Proc. of the 11th ECML PKDD, Warsaw, Poland*, p. 506–514.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2013). A personalized recommender system based on users' information in folksonomies. In *Proc. of the 22nd International Conference on World Wide Web companion*, WWW '13 Companion, p. 1215–1224.

JELASSI M. N., BEN YAHIA S. & MEPHU NGUIFO E. (2015). Towards more targeted recommendations in folksonomies. *Social Netw. Analys. Mining*, **5**(1), 68 :1–68 :18.

LANDIA N. & ANAND S. (2009). Personalised tag recommendation. *Recommender Systems & the Social Web, New York, NY, USA*, p. 83–86.

LIANG H., XU Y., LI Y. & NAYAK R. (2010). Personalized recommender system based on item taxonomy and folksonomy. In *Proceedings of the 19th ACM CIKM'10*, p. 1641–1644, New York, NY, USA : ACM.

LIPCZAK M. (2008). Tag recommendation for folksonomies oriented towards individual users. In *Proc. of the ECML/PKDD Discovery Challenge, Antwerp, Belgium*, p. 84–95.

MIKA P. (2007). Ontologies are us : A unified model of social networks and semantics. *Journal of Web Semantics.*, **5**(1), 5–15.

NANOPOULOS A., RAFAILIDIS D., SYMEONIDIS P. & MANOLOPOULOS Y. (2010). Musicbox : Personalized music recommendation based on cubic analysis of social tags. *Trans. Audio, Speech and Lang. Proc.*, **18**(2), 407–412.

QUMSIYEH R. & NG Y.-K. (2012). Predicting the ratings of multimedia items for making personalized recommendations. In *SIGIR'12*, p. 475–484, New York, NY, USA : ACM.

F. RICCI, L. ROKACH, B. SHAPIRA & P. B. KANTOR, Eds. (2011). *Recommender Systems Handbook*. Springer.

STROHMAIER M., KÖRNER C. & KERN R. (2012). Understanding why users tag : A survey of tagging motivation literature and results from an empirical study. *Web Semant.*, **17**, 1–11.

WEISS S. M. & KULIKOWSKI C. A. (1991). *Computer Systems That Learn : Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.