Améliorer la Recommandation de Ressources dans les Folksonomies par l'Utilisation de Linked Open Data

Samia Beldjoudi[1,2], Hassina Seridi[2], Abdallah Benzine[2]
[1] Ecole Préparatoire Aux Sciences et Techniques Annaba, Algérie
[2] Laboratorie de Géstion Electronique de Documents LabGED, Université Badji Mokhtar Annaba, Algérie
{beldjoudi, seridi}@labged.net
abdoulahbenzine@gmail.com

Abstract : Le Web social permet aux utilisateurs de créer, annoter, partager et rendre public les ressources qu'ils jugent intéressantes. Les folksonomies tiennent une place importante dans ces nouvelles pratiques sociales et sont utilisées dans de nombreuses applications dont les systèmes de recommandation. Aussi, l'émergence des Linked Open Data (LOD) permet d'établir des liens entre différentes entités issues de diverses sources en connectant les informations dans un unique espace de données. Dans ce travail, en plus de la prise en compte des interactions sociales afin de surmonter les problèmes d'ambigüité des tags, nous montrons comment le contenu structuré disponible à travers les LOD peut être utilisé. Les LOD sont en effet exploitées afin de pallier au manque de caractéristiques sur les ressources dans les folksonomies et faire des recommandations pertinentes et diversifiées.

Mots Clés : Folksonomies, Recommandation, Ambiguïté, Linked Open Data, Démarrage à froid, Diversité

1 Introduction

Les systèmes d'étiquetages sociaux ont gagné en popularité ces dernières années sur le Web au vu de leur simplicité pour catégoriser et retrouver les contenus en utilisant les tags. Le nombre croissant d'utilisateurs fournissant des informations sur eux-mêmes à travers leurs activités d'étiquetage sociales a induit l'émergence d'approches de profilage fondées sur les tags, qui supposent que les utilisateurs exposent leurs préférences sur des contenus au travers de tags. De ce fait, les tags peuvent être utilisés pour construire des recommandations. D'un autre côté, l'objectif principal de chaque système de recommandation est de satisfaire les intérêts des utilisateurs. L'approche classique pour cette tâche est de prédire des scores pour les ressources qui n'ont pas été évaluées par les utilisateurs et les présenter suivant l'ordre décroissant de leurs scores. Par ailleurs, ce mécanisme seul n'est généralement pas suffisant pour satisfaire les intérêts des utilisateurs. Par exemple, si un système recommande les ressources en fonction de leur popularité, il ne représente pas une plus-value significative pour l'utilisateur. En effet, les ressources ainsi recommandées à l'utilisateur ont de fortes chances d'être déjà connues par l'utilisateur puisqu'il y a de fortes chances qu'il en ait déjà entendu parler. Les critères de nouveauté et de diversité doivent être également pris en compte dans l'évaluation de la qualité d'un système de recommandation et la précision seule ne donne qu'un aperçu très partiel de l'utilité des systèmes réels.
Par ailleurs, le challenge à relever dans les systèmes de recommandation reste le problème de démarrage à froid pour les nouvelles ressources qui n'ont aucune évaluation ou pour les nouveaux utilisateurs pour lesquels le système n'a pas suffisamment d'informations.
D'autre part, les données liées (*linked data*) désignent des données suivants un paradigme fondé sur quatre règles simples : les URIs pour identifier les entités, les URLs permettant le référencement des entités, fournissant des informations utiles à ces URI fondées sur des

formats standard et la connexion et l'interconnexion à d'autres entités afin de permettre une exploration plus approfondie (Berners- Lee 2007).
Pour les données, pour être qualifiées pleinement de 'Open Linked Data' (LOD), elles devront en outre être fournies au public, disponibles sur le Web et être sous une licence ouverte. Faisant usage de formats du Web sémantique, les (LOD) mettent en œuvre la vision d'un réseau de données. Les technologies sous-jacentes permettent d'une part l'identification unique des entités via les URIs ainsi qu'une sémantique claire des relations modélisées par les liens entre les entités. Les (LOD) ont connu une croissance phénoménale au cours des dernières années. Le graphe distribué résultant des entités liées entre elles sur le web est communément appelé le nuage des (LOD) et couvre des centaines de sources de données fournissant des milliards de triplets RDF.
Ainsi, les (LOD) couvrent différents domaines allant des contenus liés aux médias, les réseaux sociaux et les contenus générés par les utilisateurs, les données bibliographiques, les sciences de la vie, la médecine, la biologie, les données géographiques, les données gouvernementales. En outre, certaines sources de données telles que DBPedia fournissent des informations générales, inter-domaines et ainsi jouent un rôle clé dans la connexion des données provenant des domaines très différents.
Dans ce papier, nous considérons le domaine du Web social sémantique et particulièrement les problèmes liés à la recommandation de ressources dans les folksonomies. Nous proposons une méthode pour analyser les profiles des utilisateurs selon leurs activités d'étiquetage afin d'améliorer la recommandation des ressources. L'efficacité de la recommandation dépend de la résolution des problèmes inhérents aux folksonmies. Dans notre processus de recommandation, nous montrons comment le problème de l'ambigüité peut être réduit en tenant compte des similarités sociales calculées sur les folksonomies combinées avec les similarités entre ressources dans les LOD. Nous utilisons également la force des LOD pour diversifier la recommandation dans les systèmes d'étiquetages sociaux et ce par l'exploration d'entités inter-liés.
Ce papier est organisé comme suit : la section 2 est un survol des travaux connexes. La section 3 est dédiée à la présentation de l'approche. Dans la section 4, les résultats sur les expérimentations sont présentées et discutées pour conclure sur les performances de notre approche. Les conclusions et les perspectives sont décrites dans la section 5.

2 Travaux connexes

Beaucoup de recherches dans le passé ont proposé d'utiliser les ontologies et les taxonomies pour améliorer la qualité des systèmes de recommandation conventionnels (Maidel et al, 2008, Middleton et al., 2004, Anand et al., 2007). Ces dernières années, avec l'émergence des LOD, une nouvelle classe de systèmes de recommandation a vu le jour nommée systèmes de recommandation basés sur les LOD. La communauté du Web sémantique et des systèmes de recommandation s'intéressent de plus en plus à cette nouvelle topologie de systèmes de recommandation. La plus part des travaux liés à ces thématiques ont essayé de réutiliser et d'adapter quelques idées issues des systèmes de recommandation ontologiques aux LOD en s'adaptant à leurs caractéristiques propres alors que d'autres ont proposé de nouvelles approches conçues spécifiquement pour les technologies des Linked Open Data et ont alors proposé de nouvelles applications des systèmes de recommandation pour celles-ci. Dans ce qui suit, nous allons passer en revue les contributions les plus significatives. L'une des approches qui exploite les Linked Open Data pour construire des systèmes de recommandation est celle de (Marie et al, 2013) dans laquelle des datasets des LOD sont utilisés pour une exploration personnalisée utilisant une méthode d'activation en diffusion. Une méthode d'activation en diffusion a été utilisée afin de trouver des relations sémantiques

entre des items appartenant à différents domaines. Un système de recommandation entièrement basé sur SPARQL nommé RecSPARQL a été présenté dans (Ayala et al, 2014). L'outil proposé étend la syntaxe et la sémantique de SPARQL afin de permettre un filtrage collaboratif flexible et générique et une recommandation basée contenu sur des graphes RDF. Dans (Khrouf et Troncy, 2013), les auteurs présentent un système de recommandation événementiel basé sur les Linked Data et la diversité des utilisateurs. Une extension sémantique du modèle SVD+++ nommé SemanticSVD+++ est présentée dans (Rowe, 2014). Elle intègre des catégories sémantiques d'items dans le modèle. Ce modèle est également capable de considérer l'évolution au fil du temps des préférences des utilisateurs. Dans (Rowe, 2014), les auteurs améliorent le travail précédent pour tenir compte des items démarrant à froid. Ils introduisent des sommets-noyaux afin d'obtenir des informations sur les catégories sémantiques non évaluées en démarrant des catégories connues. Enfin, dans (Dojchinovski et Vitvar, 2014) les auteurs proposent l'utilisation de techniques de recommandation afin de fournir un accès personnalisé aux Linked Data. La méthode de recommandation proposée est un système de filtrage collaboratif utilisateur-utilisateur où la similarité entre les utilisateurs prend en compte les points communs et l'informativité des ressources au lieu de les considérer comme de simples identificateurs.

D'autre part, dans les systèmes d'étiquetage social, l'objectif général de la recommandation de ressources est d'assurer la quantité et la pertinence des ressources recommandées. Parmi les travaux traitant de ce problème, nous pouvons citer (Huang et al, 2011) qui a proposé un système de recommandation qui utilise les préférences les plus récemment identifiées dans les tags des utilisateurs. (Zanardi et al, 2011) ont proposé une méthode destinée à étendre les capacités de recherche des collections digitales ciblant les domaines académiques et scolaires. (Beldjoudi et al., 2011, 2012) ont proposé une méthode pour analyser les profiles des utilisateurs afin d'améliorer la recommandation de ressources dans les folksonomies. L'objectif est d'enrichir les profils des utilisateurs avec des ressources pertinentes en résolvant le problème de l'ambigüité des tags durant la recommandation.

Le problème de la diversité des résultats a déjà été traité dans la Recherche d'Information(RI) mais sous un angle diffèrent. Ce problème est traité par la (RI) afin de résoudre celui de l'ambigüité et/ou de la sous-spécification des requêtes des utilisateurs. Dans la Recherche d'Information, l'accent est mis sur l'élargissement des items recommandés présentés à l'utilisateur (diversité) et la promotion d'items moins connus (nouveauté) ou d'items non familiers pour un utilisateur donné. Quelques recherches ont été menées dans ce terrain, et ont connu un intérêt croissant à cause de l'importance de la diversité et de la nouveauté dans la communauté des systèmes de recommandation. Néanmoins, il reste encore un espace de recherche considérable dans l'amélioration de la recommandation de ressources dans les systèmes d'étiquetage social par l'utilisation des Linked Open Data afin d'assurer des résultats précis et diversifiés.

3 Description de l'approche

Une folksonomie est définie comme un modèle triparti où les ressources Web sont associées à un utilisateur par une liste de tags. Formellement, une folksonomie est un tuple F = < U, T, R, A> où U, T et R représentent respectivement un ensemble d'utilisateurs, un ensemble de tags et un ensemble de ressources et A représente les relations entre les trois éléments précédents, c'est-à-dire $A \subseteq U \times T \times R$ (Mika, 2005)..

Nous extrayons trois réseaux sociaux à partir d'une folksonomie, qui représentent trois points de vue différents sur les interactions sociales: un réseau relatif aux tags et aux utilisateurs et un second concernant les tags et les ressources et un troisième concernant les utilisateurs et les ressources. Nous représentons ces réseaux sociaux par trois matrices TU, TR, UR :

-TU = $[X_{ij}]$ où : $X_{ij} = \begin{cases} 1 \text{ si } \exists \text{ r} \in \text{R}, < uj, ti, r > \in \text{A} \\ 0 \text{ autrement} \end{cases}$

-TR = $[Y_{ij}]$ où: $Y_{ij} = \begin{cases} 1 \text{ si } \exists \text{ u} \in \text{U}, < u, ti, rj > \in \text{A} \\ 0 \text{ autrement} \end{cases}$

-UR = $[Z_{ij}]$ où: $Z_{ij} = \begin{cases} 1 \text{ si } \exists \text{ t} \in \text{T}, < ui, t, rj > \in \text{A} \\ 0 \text{ autrement} \end{cases}$, RU, RT et UT sont transposées dans les matrices UR, TR and TU.

C'est ce qui nous permet d'analyser les corrélations issues des différentes interactions sociales. Nous utilisons Pajek, un outil qui a déjà été utilisé par Mika pour analyser les grands réseaux (Mika, 2005).

Dans cet article, nous proposons une méthode pour analyser les profils des utilisateurs d'après leurs tags afin de trouver des ressources intéressantes et les recommander. L'objectif est d'enrichir les profils des utilisateurs de folksonomies avec des ressources pertinentes. Nous supposons que le partage automatique de ressources renforce les liens sociaux entre les acteurs et nous exploitons cette idée afin de réduire l'ambigüité des tags dans le processus de recommandation en augmentant le poids associé aux ressources web selon les similarités sociales. Nous nous sommes basés sur des règles d'association qui sont une méthode puissante pour découvrir des corrélations intéressantes entre un grand ensemble de données sur le Web. Pour appliquer une méthode de règle d'association dans les folksonomies, nous avons représenté chaque utilisateur dans la folksonomie par un ID de transaction et les tags qu'ils utilisent par l'ensemble des éléments qui sont dans cette transaction (Beldjoudi et al., 2012).

Notre objectif est de trouver des corrélations entre les balises, c.à.d. de trouver des tags apparaissant fréquemment ensembles afin d'en extraire ceux qui ne sont pas utilisés par un utilisateur particulier, mais qui sont souvent utilisés par d'autres utilisateurs proches de lui. Par exemple, considérons un ensemble de données dans lequel il s'apparait que de nombreux utilisateurs utilisant le tag *Software* utilisent également le tag *Java*. Nous cherchons à extraire une règle *Software* → *Java* afin que nous puissions enrichir les profils des utilisateurs qui emploient le tag *Software*, mais pas le tag *Java*, par les ressources taggués avec *Java*. Une fois que les règles sont extraites, notre système de recommandation se déroule comme suit: Pour chaque règle extraite, nous testons si les balises qui sont dans l'antécédent de la règle sont utilisées par l'utilisateur actuel. Si tel est le cas, alors les ressources taggués avec chaque balise trouvée dans le conséquent de la règle sont candidates à être recommandée par le système.

L'efficacité de la recommandation dépend de la résolution des problèmes inhérents aux folksonomies. Dans notre approche, nous abordons les problèmes d'ambiguïté des tags, les variations orthographiques (ou synonymie) et le manque de liens sémantiques entre tags. Le détail sera décrit dans les prochains paragraphes.

3.1 Exploiter les similarités sociales et le LOD pour surmonter l'ambiguïté des tags et le démarrage à froid lors de la recommandation

Selon (Mathes, 2004), «Les problèmes inhérents à un vocabulaire non contrôlé conduit à un certain nombre de limites et les faiblesses dans les folksonomies. L'ambiguïté des tags peut être levée lorsque les utilisateurs appliquent le même tag de différentes manières".

Une balise peut avoir plusieurs significations, c.à.d. référer à plusieurs concepts. Par conséquent, un système de recommandation basé sur les tags recommande aussi bien des ressources relatives aux fruits ou aux ordinateurs à un utilisateur qui recherche avec le tag "apple". La résolution de problème d'ambiguïté est particulièrement cruciale dans notre approche, où certaines balises qui sont utilisées pour recommander des ressources ne sont pas utilisées directement par l'utilisateur mais déduit des règles d'association (Beldjoudi et al, 2011). Pour résoudre le problème d'ambiguïté lors de la recommandation, nous proposons de

mesurer la similarité entre utilisateurs afin d'identifier ceux qui ont des préférences similaires et par conséquent adapter la recommandation aux profils d'utilisateurs (voir Algorithme 1).
Nous expliquons comment les similarités sociales et les LOD sont utilisés pour surmonter l'ambiguïté des tags et le problème de démarrage à froid lors de la recommandation dans ces deux étapes :
- *Première étape*: Pour chaque règle d'association A → B dont l'antécédent s'applique à un utilisateur actif u_x , nous mesurons les similarités entre cet utilisateur et les utilisateurs qui utilisent les tags qui se trouvent dans le conséquent de la règle. Les ressources associées à ces tags sont recommandées à cet utilisateur en fonction de ces similarités. Pour mesurer la similarité entre deux utilisateurs u1 et u2, les deux sont représentés par un vecteur binaire représentant tout leurs tags (extrait de la matrice UT) et on calcule le cosinus de l'angle entre les deux vecteurs: $\text{sim}(u_1, u_2) = cos(v_1, v_2) = \frac{v_1.v_2}{\|v_1\|^2\|v_2\|^2}$ (1)

Selon (Cattuto et al., 2008) et (Koerner et al., 2010), le calcul de similarité avec la formule cosinus donne de bons résultats en un coût de calcul très raisonnable, car il a une complexité linéaire. Nous insistons sur le fait que la distribution des tags en fonction des ressources et des utilisateurs dans les folksonomies suit une loi de calcul de puissance: la plupart des ressources sont marquées par un petit nombre d'utilisateurs, et de nombreux tags ne sont utilisés que par quelques utilisateurs, une propriété qui conduit à une faible valeur de r (le nombre de ressources dans la matrice RU) et n (le nombre d'utilisateurs dans la matrice UT). Par conséquent, notre approche peut évoluer dans les très grandes bases de données.
- *Deuxième étape:* Pour éviter le problème de démarrage à froid qui résulte généralement du manque de données requises par le système afin de faire une bonne recommandation, lorsque l'utilisateur du système de recommandation n'est pas encore semblable à d'autres utilisateurs, nous proposons d'exploiter les liens sémantiques entre les ressources dans le LOD. Celles-ci peuvent être considérées comme une source fiable et riche d'informations. Elles aident les systèmes de recommandation à résoudre certains problèmes, tels que le problème du démarrage à froid et l'analyse de contenu limité. On se fonde pour cela sur une mesure robuste des similarités entre les ressources en utilisant les LOD. Dans cette approche, nous utilisons le Open Linked Data pour apprécier la similarité entre les ressources d'une folksonomie en utilisant leurs ressources correspondantes sur les LOD (Fig1) (c.à.d. nous mesurons la similarité entre les ressources qui seraient recommandées par le système (celles qui sont liées à un tag apparaissant dans la conséquence d'une règle d'association) et celles qui sont déjà recommandées à l'utilisateur. La similarité entre deux ressources est calculée en utilisant l'indice de Jaccard défini comme suit: $sim(R1, R2) = J(R1, R2) = \frac{|R1 \cap R2|}{|R1 \cup R2|}$ (2)

Chaque ressource Rx est définie par ses caractéristiques c.à.d. des triplets de type (Rx, prédicat, Ry), où prédicat indique le type de la relation et Ry représente le nœud cible (c.à.d. le nœud connecté à l'autre extrémité de la relation).

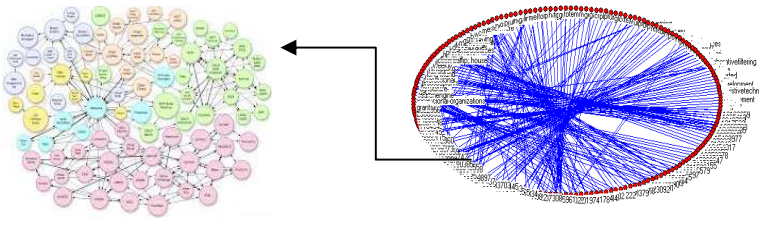

FIGURE 1 -Lier les ressources dans la base del.icio.us (qui sont représentées par l'outil Pajek) à leurs ressources correspondantes dans DBpedia

Algorithme1 : Recommandation personnalisée de ressources

Entrée: Une folksonomie : F< U, T, R, A>, *S1, S2*: des nombres entiers positifs
Sortie: L_r: liste de ressources à recommandées

Début

1. Génération de N Règles associative $\{t_A \rightarrow t_B\}$

2. Pour K=1 jusqu'à N faire,

3. Construire la matrice $UR|t_B = [Z_{ij}]$ où: $Z_{ij} = \begin{cases} 1 \text{ if } < u_i, t_B, r_j > \in A \\ 0 \text{ otherwise} \end{cases}$

4. Construire la matrice $UT|r_j = [X_{ij}]$ où : $X_{ij} = \begin{cases} 1 \text{ if } < u_x, ti, r_j > \in A \\ 0 \text{ otherwise} \end{cases}$

5. Construire la matrice UU= UT*TU

6. Calculer Sim-u= Cos (v1,v2)

7. Si Sim-u >= S1 alors, $L_r = L_r \cup \{rj\}$

8. Sinon Calculer Sim-r = $J(rj, rm)$ en utilisant LOD

9. Si Sim-r >= S2 L=L ∪ {rj}

10. Fin Si

11. Fin Sinon

12. Fin Si

13. Fin Pour

14. Renvoyer (L_r);

Fin

3.2 Assurer la diversité dans la Recommandation

Lors de l'utilisation d'un système de recommandation tels que Amazon.com, Netflix, etc. on peut rencontrer le problème suivant: si un profil d'utilisateur est composé d'un couple de livres de "Victor Hugo", un moteur de recommandation axée uniquement sur la précision peut fournir une liste composée principalement d'autres livres de "Victor Hugo". Bien qu'il soit très probable que l'utilisateur va aimer les livres recommandés, il est clair que la recommandation n'est pas très utile dans le sens de:

-Le Manque de diversité, probablement un plus petit échantillon de livres de "Victor Hugo" aurait été aussi utile pour découvrir le travail de l'auteur et aurait donné l'espace pour d'autres livres intéressants pour d'autres auteurs; et

-Le Manque de nouveauté, puisque "Victor Hugo" est un auteur très connu, pour lequel un système de recommandation n'est même pas nécessaire.

Cette situation ouvre deux questions: Pourquoi le système fournit de tels résultats ? Comment résoudre ce problème ? En règle générale, les systèmes de recommandation sont entrainés à minimiser l'erreur de prédiction, de sorte que des aspects tels que la redondance et l'évidence ne sont généralement pas considérés. Un autre problème réside dans la sous-spécification du profil d'utilisateur. Comme il contient qu'un livre d'un auteur unique, une approche de filtrage collaboratif pur est susceptible de trouver la plupart des connexions à d'autres utilisateurs qui auront plus de livres du même auteur. Enfin, même si l'utilisateur avait acheté ou naviguer vers des livres d'autres auteurs, ceux de "Victor Hugo" resteront toujours populaires et donc seront inévitablement favorisés par un algorithme de recommandation standard.

Pour résoudre ce dilemme dans les folksonomies, nous proposons d'extraire à partir des ressources les caractéristiques les plus populaires trouvés dans le profil de l'utilisateur (c.à.d. les caractéristiques qui intéressent l'utilisateur au moment de choisir ses ressources) et ensuite explorer le graphe de LOD afin d'en extraire des ressources liées à ces caractéristiques. Par exemple, considérons le cas suivant:

FIGURE 2 - Le profil de l'utilisateur contenant un ensemble de films de Leonardo Dicaprio

Dans cet exemple, le profil de l'utilisateur est composé des ressources (R1, R2, R3, R4 et R5), dont l'intersection entre les caractéristiques de ces ressources doit être calculé (R1∩ R2 R3 ∩

∩R4 ∩ R5). On extrait ainsi les caractéristiques les plus populaires qui intéressent l'utilisateur quand il choisit de tagguer ses propres ressources. Ensuite, pour chaque caractéristique (Pi) dans le résultat de l'intersection, nous allons explorer le graphe LOD à trois niveaux pour extraire d'autres ressources (R6) ayant cette caractéristique ou ayant un lien direct / indirect avec ces dernières (R7, R8 resp) .Nous avons fixé le niveau d'exploration à 3 afin d'éviter de biaiser les résultats de recommandation.
Nous avons dans l'exemple ci-dessus (R1∩ R2 ∩ R3 ∩R4 ∩ R5) = {Leonardo Dicaprio ...}. En explorant le sous-graphe suivant, nous constatons que la ressource "Leonardo DiCaprio" est liée à la ressource "OSCARS" via le prédicat (has). A son tour, la ressource "OSCARS" est lié via le prédicat (winner) avec la ressource "Eddie Redmayne". Par conséquent, nous pouvons recommander des films de « Eddie Redmayne » par exemple à l'utilisateur actuel.

FIGURE 3 – Un sous graphe de LOD

Avec cette méthode, nous nous assurons que la liste des ressources à recommander est diversifiée, où chaque utilisateur peut obtenir autres ressources différentes à celles qui se trouvent dans son profile, même si elles ne figurent pas dans les profils de ses voisins dans le réseau social.
Chaque ressource recommandée par le système est d'abord associée un poids initial basé sur les similarités entre utilisateurs. Au-dessus d'un seuil fixé dans [0..1], nous qualifions la ressource comme fortement recommandée. Sous ce seuil, nous considérons la similarité entre ressources et nous recommandons fortement de même les ressources que les poids calculés sur LOD sont au-dessus d'un seuil donné. Nous notons que notre système de recommandation est flexible, puisque l'utilisateur peut interagir pour accepter ou rejeter les ressources recommandées.

3 Les résultats expérimentaux

Nous avons montré dans la section précédente que notre approche utilise les dimensions sociales et les sémantiques du web afin d'améliorer le processus de recommandation. Cette section donne des détails sur l'implémentation pour permettre son évaluation et montrer son efficacité.
Afin de valider notre approche, nous avons conduit une expérimentation avec la base del.icio.us. Notre base de test comprend 58588 assignations de tags impliquant 12780 utilisateurs, 30500 tags parmi lesquels certains sont ambigus et ont des orthographes différentes et 14390 ressources chacune étant associée à plusieurs tags et à plusieurs utilisateurs. Notre système a extrait un ensemble de 946 règles d'association de la base de données avec un support égal à 0.5 et une confiance égale à 0.6.
La principale base de données LOD utilisée pour nos expérimentation est DBPedia, l'une des initiatives du web sémantique ayant eu le plus de succès. Afin d'évaluer notre système de

recommandation basé sur les LOD, les ressources de la base del.cio.us doivent être mise en correspondance avec celles de DBpedia.
Les LOD peuvent être interrogées au travers de leurs endpoints SPARQL. Pour DBpedia, cela permet à n'importe qui d'effectuer des requêtes complexes sur n'importe quel sujet disponible dans Wikipedia. Par exemple, on peut savoir simplement quels acteurs ont joué dans le film « Le Revenant » via la reqête SPARQL :
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
SELECT ? actor WHERE {dbpedia:the_revenant dbpedia-owl:starring ?actor .}
Cet exemple permet de voir comment on peut extraire de nombres informations aussi bien sur une ressource spécifique que sur plusieurs d'entre elles. A partir de l'url associée à un item, il est possible d'extraire le sous-graphe associé en effectuant plusieurs requêtes SPARQL en utilisant une stratégie profondeur d'abord à profondeur limitée.
La sémantique des classes LOD et leurs relations sont décrites grâce à des ontologies. Par exemple, la ressource dbpedia :Leonard-dicaprio dans DBpedia est une instance de la classe dbpedia-owl :Person qui est à son tour une sous-classe de dbpedia-owl :Agent. La sémantique des propriétés est également définie dans une telle ontologie. Par exemple, la propriété dbpedia-owl :starring qui relie dbpedia :The_revenant à dbpedia :Leonardo-dicaprio a pour domaine dbpedia-owl :Workand et pour *range* dbpedia-owl :Actof qui est une sous-classe de dbpedia-owl :Person.

4.1 Le protocole expérimental

Idéalement, dans le but d'évaluer la qualité d'un système de recommandation, nous devons montrer que les ressources recommandées sont réellement acceptées par l'utilisateur. Mais pour le savoir, nous devrions interroger les utilisateurs des bases de données choisies et leur demander s'ils ont apprécié l'ensemble des ressources proposées. Puisque ceci est impossible, nous avons retiré aléatoirement certaines ressources du profil de chaque utilisateur et nous avons appliqué notre approche sur l'ensemble des données restantes afin de voir si les ressources retirées sont recommandées à leurs utilisateurs respectifs ou pas. Si elles sont recommandées, nous pouvons alors conclure que le système a correctement estimé les préférences de l'utilisateur. Afin de tester les performances de notre approche, nous avons suivi les étapes suivantes :

a) Evaluation de la capacité à dépasser le problème de l'ambigüité

Pour atteindre ce but, nous avons commencé par sélectionner un ensemble de 1154 tags ambigus de la base del.ici.us. Nous avons alors aléatoirement retiré des ensembles de ressources correspondants à ces tags ambigus. Ce processus a été répété cinq fois pour chaque tag dans le but d'effectuer une cross-validation. En d'autres termes, pour chaque tag, nous avons aléatoirement divisé l'ensemble des ressources correspondantes en cinq parties et nous avons ensuite sélectionné la partie à retirer dans chaque évaluation pour l'utiliser comme un ensemble de test. Ce processus a été répété cinq fois et à chaque nous avons choisi un ensemble de test diffèrent.

- Résultats expérimentaux : Pour évaluer la qualité de notre système de recommandation, nous avons utilisé trois métriques : rappel, précision et F1 qui est une combinaison des deux premières. 107 règles d'association ont été extraites avec un support égal à 0.5 et une confiance égale à 0.6. Ensuite, les trois métriques ont été calculées pour chaque participant. La table 1 présente les valeurs moyennes des métriques :

Table 1: Précision, rappel et F1 moyennes des recommandations

Précision	Rappel	F1
77%	83%	80%

Ces résultats montrent que, en appliquant les règles d'association extraites, les ressources associées à des tags non ambigus sont très recommandées. Cela montre également que, dans le cas des règles faisant intervenir des tags ambigus, notre système recommande à l'utilisateur des ressources qui sont proches de ses intérêts avec un haut niveau de recommandation et celles qui sont éloignées de ses intérêts avec un faible niveau de recommandation.

b) **Evaluation de la capacité à dépasser le problème de variations d'orthographes**

Pour atteindre ce second objectif, nous avons commencé par sélectionner un ensemble de tags contenant des termes avec beaucoup d'orthographes différentes. Cela donne un ensemble de 2417 tags extraits de la base del.icio.us. Ensuite, nous avons aléatoirement retiré des ressources étiquetées par ces tags afin de déterminer si le système les recommande aux bons utilisateurs. Ce processus a été répété cinq fois afin d'effectuer une validation croisée.

- Résultats expérimentaux : Basé sur nos ensembles de test, 127 règles d'association ont été extraites et cela avec un support égal à 0.5 et une confiance égale à 0.6. Ensuite, nous calculé les mêmes métriques que précédemment pour chaque utilisateur. La table 2 les valeurs moyennes obtenues pour chaque métrique :

Table 2 : Précision, rappel et F1 moyennes des recommandations

Précision	Rappel	F1
69%	80%	75%

4.2 Discussion

Nous pouvons conclure de l'analyse des résultats précédents que, dans tous les cas, la précision, le rappel et la métrique F1 de notre approche sont très prometteuses pour la base del.icio.us. Ces résultats indiquent que l'utilisation de règles d'association et des similitudes sociales combinées aux LOD permet de tenir compte du profil de l'utilisateur lors de la recommandation de ressources. En effet, ces résultats montrent que notre approche réussit à distinguer entre les tags ambigus et permet de tenir en compte les variations de l'orthographe durant la recommandation de ressources. La table 3 présente l'écart-type de la précision, du rappel et de la métrique F1 dans la base del.icio.us pour l'ambigüité des tags et le problème des orthographes multiples.

Table 3: L'écart-type des trios métriques pour l'ambigüité des tags et les orthographes multiples

	Précision	Rappel	F1
L'ambiguité des tags	5%	6%	5%
Orthographes multiples	8%	5%	4%

Dans les deux cas, ces valeurs sont très petites ce qui indique que les valeurs de ces trois mesures pour chaque utilisateur sont très proches de la moyenne. Les valeurs moyennes (présentés dans les tables 1 et 2) étant très prometteuses pour la communauté en général, les petites valeurs de l'écart-type indiquent que les métriques sont également très prometteuses pour chaque utilisateur.

4.3 Le Choix de la valeur optimale pour le support et la confiance

L'objectif de la fouille par les règles d'association est de trouver toutes les règles qui satisfont un support minimum et des restrictions de confiance. Plus on augmente la valeur du support, plus les règles extraites sont évidentes et alors moins elles sont utiles pour l'utilisateur. Il en résulte qu'il est nécessaire de choisir une valeur pour le support suffisamment basse afin d'extraire une information importante. Malheureusement, lorsque le seuil du support est trop bas, la quantité de règles extraites devient très grande rendant l'analyse de ces règles difficile. La confiance est une estimation de la précision des règles dans le futur. Cela représente la confiance désirée dans les règles.

Une certaine expertise est nécessaire afin de trouver les bonnes valeurs du support et de la confiance qui permettront d'obtenir les meilleures règles qui impactent la métrique F1. Pour trouver les valeurs optimales de ces deux paramètres, deux expérimentations ont été menées. Dans la première expérience, la valeur optimale du support a été recherchée en utilisant la

base del.icio.us. Nous avons fait varier le support sur un intervalle de 0.1 à 1 et nous avons sélectionné la valeur donnant les meilleures performances. La figure 4 montre l'évolution de la métrique F1 par rapport à la valeur du support.

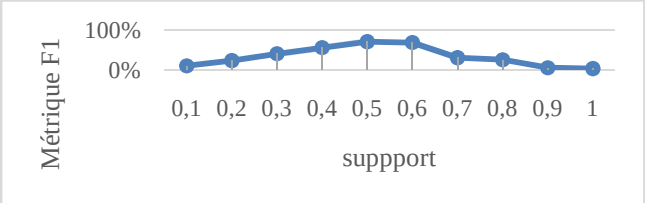

Figure 4 – Valeur optimale du support

Comme nous pouvons le voir sur cette figure, la meilleure valeur du support qui produit la plus grande valeur de la métrique F1 est 0.5. La seconde expérience concerne la recherche de la valeur optimale de la confiance en utilisant également la base del.icio.us pour un support minimal égal à 0.5. On a fait varier la confiance de la même façon que le support. La figure 5 montre l'évolution de la valeur de la métrique F1 par rapport à la confiance.

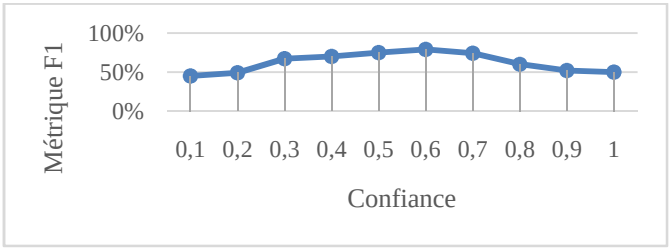

Figure 5 – Valeur optimal de la confiance

On en déduit que la valeur optimale de la confiance est 0.6. Il en résulte que des valeurs appropriées pour le support et la confiance sont respectivement 0.5 et 0.6.

4.4 Diversité dans la recommandation

Afin d'évaluer l'efficacité de notre approche pour donner des recommandations diversifiées, la métrique de diversité Intra-List proposée par (Ziegler et al., 09) est utilisée**.**

Dans cette section, nous évaluons la diversité de notre approche de recommandation et la comparons à la précision de celle-ci. Nous voulons ainsi voir si une augmentation de la diversité a bien lieu grâce à notre approche et si celle-ci a un impact sur la précision des recommandations. Nous avons testé notre approche sur trois niveaux d'exploration de LOD. Le nombre d'utilisateurs est égal à 20.

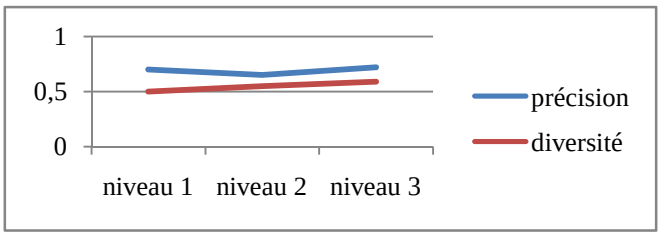

FIGURE 6 **–** Diversité Vs. Performance de la Recommandation

Les résultats présentés dans la figure 6 montrent que la recommandation basée sur les LOD augmente le taux de diversité avec le nombre d'items à recommander, provoquant juste une légère perte de précision.

La diversité dans notre approche est encore en cours d'évaluation et les premiers résultats montrent l'utilité d'explorer les LOD pour augmenter la diversité lors de la recommandation de ressources.

4.5 Passage à l'échelle

Les systèmes de recommandation étant destinés à aider les utilisateurs à naviguer dans de larges collections d'items, l'un de nos objectifs est de passer à l'échelle des Datasets réels. Il est donc important de mesurer la vitesse avec laquelle notre approche fournit des recommandations. Dans cette sous-section, on discutera de l'impact qu'à l'augmentation du nombre d'utilisateurs sur le temps d'exécution de notre approche. Afin de montrer la scalabilité de notre approche, nous avons mesuré le temps d'exécution requis afin de faire des recommandations dans la base del.icio.us avec un nombre d'utilisateurs allant de 1000 à 11500. La figure 7 montre que le temps d'exécution (en secondes) croît linéairement avec l'augmentation de la taille de la base de données. Cela signifie que notre approche répond bien à ce problème puisque l'augmentation du nombre d'utilisateurs dans la base de données provoquera approximativement une augmentation linéaire du temps d'exécution.

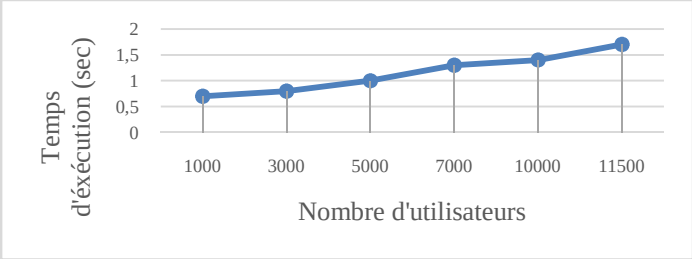

FIGURE 7 – Performance de notre approche lorsque la taille la base de données croît

5 Conclusion

Dans cette contribution, nous avons exploité la force de l'aspect social des folksonomies afin de permettre à chaque utilisateur de la communauté de bénéficier des ressources taguées par ses voisins dans le réseau social basé sur la recommandation de ressources. Nous avons vu l'importance d'analyser le profil de l'utilisateur afin de réaliser une recommandation dynamique et par conséquent l'importance de venir à bout des problèmes sémantiques inhérents aux folksonomies durant la recommandation. La méthode suivie est basée sur la similarité entre les utilisateurs dans certains cas et entre ressources LOD dans d'autres cas afin de venir à bout du problème du démarrage à froid lors de la recommandation. Les premiers résultats montrent l'intérêt d'explorer le graphe des LOD afin d'assurer la diversité lors de la recommandation de ressources personnalisées dans les systèmes d'étiquetage social.

Références

S. S. Anand, P. Kearney, et M. Shapcott. Generating semantically enriched user profiles for web personalization. ACM Trans. Internet Technol., 7(4), Oct. 2007.

V. A. A. Ayala, M. Przyjaciel-Zablocki, T. Hornung, A. Schatzle, et G. Lausen. Extending sparql for recommendations. In Proceedings of Semantic Web Information Management on Semantic Web Information Management, SWIM'14, pages 1:1-1:8, New York, NY, USA, 2014. ACM.

S. Beldjoudi, H. Seridi et C. Faron-Zucker. Ambiguity in Tagging and the Community Effect in Researching Relevant Resources in Folksonomies. In Proc. of ESWC workshop User Profile Data on the Social Semantic Web, 2011.

S. Beldjoudi, H. Seridi et C. Faron-Zucker. Improving Tag-based Resource Recommendation with Association Rules on Folksonomies. In Proc. Of ISWC workshop on Semantic Personalized Information Management: Retrieval and Recommendation, 2011.
S. Beldjoudi, H. Seridi et C. Faron-Zucker. Personalizing and Improving Tag-based Search in Folksonomies. In Proc. Of the 15th International Conference on Artificial Intelligence Methodology, Systems, Applications (AIMSA), Springer LNAI 7557, pp. 112-118, 2012.
P. De Meo, G. Quattrone, et D. Ursino. 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, 2010.
M. Dojchinovski et T. Vitvar. Personalised access to linked data. In EKAW, pages 121-136, 2014.
C.L Huang, H.Y Chien, et M Conyette, Folksonomy-based Recommender Systems with User.s Recent Preferences, World Academy of Science,Engineering and Technology 78, 2011.
H. Khrouf et R. Troncy. Hybrid event recommendation using linked data and user diversity. In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, pages 185-192, New York, NY, USA, 2013.
N. Marie, O. Corby, F. Gandon, et M. Ribiere. Composite interests' exploration thanks to on-they linked data spreading activation. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, HT '13, pages 31-40, New York, NY, USA, 2013.
S. E. Middleton, N. R. Shadbolt, et D. C. De Roure. Ontological user profiling in recommender systems. ACM Trans. Inf. Syst., 22:54-88, January, 2004.
P. Mika. Ontologies are us: A unified model of social networks and semantics. In Proc. of 4th Int. Semantic Web Conference (ISWC 2005), Galway, Ireland, volume 3729 of LNCS. Springer, 2005.
B. Mobasher, X. Jin, et Y. Zhou. Semantically enhanced collaborative filtering on the web. In B. Berendt, A. Hotho, D. Mladenic, M. Someren, M. Spiliopoulou, et G. Stumme, editors, Web Mining: From Web to Semantic Web, volume 3209 of Lecture Notes in Computer Science, pages 57-76. Springer Berlin Heidelberg, 2004.
M. Rowe. Semanticsvd++: incorporating semantic taste evolution for predicting ratings. In 2014 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2014, 2014.
M. Rowe. Transferring semantic categories with vertex kernels: recommendations with semanticsvd++. In The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, 2014.
V. Zanardi, L. Capra. A Scalable Tag-based Recommender System for New Users of the Social Web. In: Proc. of the 2nd International Conference on Database and Expert Systems Applications, 2011.
C. Ziegler, G. Lausen et G. Georges-Köhler-Allee. Making Product Recommendations More Diverse. IEEE Data Eng. Bull., 32(4), pp. 23-32, 2009.