Un cycle de vie complet pour l'enrichissement sémantique des folksonomies

Freddy Limpens, Fabien Gandon, Michel Buffa

Un cycle de vie complet pour l'enrichissement sémantique des folksonomies

Freddy Limpens*, Fabien Gandon*
Michel Buffa**

*EPI Edelweiss - INRIA Sophia Antipolis
Freddy.Limpens, Fabien.Gandon@inria.fr,
http://www-sop.inria.fr/edelweiss/
**I3S - KEWI, UNSA - CNRS
buffa@unice.fr
http://www.i3s.unice.fr/I3S/labos/labo2.html

Résumé. Les tags fournis par les utilisateurs des plateformes de tagging social ne sont pas explicitement liés sémantiquement, et ceci limite considérablement les possibilités d'exploitation de ces données. Nous présentons dans cet article notre approche pour l'enrichissement sémantiques des folksonomies qui intègre une combinaison de traitements automatiques ainsi que la capture des contributions de structuration des utilisateurs via une interface ergonomique. De plus, notre modèle supporte les points de vue qui divergent tout en permettant de les combiner en respectant leur cohérence locale. Cette approche s'adresse aux communautés de connaissances collaborant en ligne, et en intégrant leurs usages, nous sommes en mesure de proposer un cycle de vie complet pour le processus de structuration sémantique des folksonomies. La navigation dans les données de tagging est ainsi améliorée, et les folksonomies peuvent alors être directement intégrées dans la construction de thesauri.

1 Introduction

Le tagging social est un moyen simple et populaire de classification par mot clé. L'ensemble des mots utilisés forme une folksonomie que l'on peut considérer comme "le vocabulaire de la communauté". Cependant les folksonomies résultant de cette pratique ne possèdent aucune structure car les tags ne sont pas reliés sémantiquement et manquent de précision, notamment à cause des variations orthographiques, ce qui limite leur potentiels d'exploitation pour la navigation dans les corpus tagués. Nous présentons dans cet article notre approche de l'enrichissement sémantique de folksonomie. Notre contribution consiste en un cycle complet du processus qui combine traitements automatiques et contributions des utilisateurs.

Un exemple-type de nos communautés cibles est représenté par l'Ademe [1] qui cherche à améliorer la structuration et l'indexation des sources d'information utilisées dans les processus de veille scientifique et technique. Dans ce scénario, nous pouvons distinguer trois types

1. Agence De l'Environnement et de la Maitrise de l'Energie http ://www.ademe.fr

d'utilisateurs : (1) les ingénieurs experts spécialistes d'un domaine, (2) les documentalistes en charge de l'indexation du corpus, et (3) le public ayant accès à une partie des resources Ademe via son site Web. Les documentalistes cherchent à structurer une base de mots-clés, assimilable à une folksonomie contrôlée, pour la rapprocher de la structure d'un thésaurus. La difficulté réside ici dans la nécessité d'inclure les différents points de vue pouvant émerger en interne entre les experts ou les membres du public.

Notre travail s'est principalement intéressé à proposer des moyens d'enrichir sémantiquement cette folksonomie contrôlée tout en préservant le support de points de vues divergents (quant à la structuration) tout au long de son cycle de vie. Ceci permet à chaque utilisateur de maintenir sa propre structuration des tags tout en contribuant à un point de vue global qui est généré a posteriori par un systèmes détectant et solutionnant les éventuels conflits entre points de vue et maintenu par un utilisateur référent. Nous proposons donc une solution au problème du goulet d'étranglement à l'acquisition de connaissances en permettant d'inclure un maximum d'expertise dans le processus de structuration de folksonomie.

Cette article s'organise de la manière suivante. En section deux nous donnons un apperçu des recherches visant à rapprocher folksonomies et representations de connaissances structurées. La section trois présente le cycle de structuration de folksonomie que nous proposons. La section quatre présentera les méthodes de traitements automatiques, et la section cinq la capture et la combinaison des contributions individuelles visant à fournir une folksonomie semantiquement structurée et multi-points de vue.

2 Etat de l'art et positionnement

L'enrichissement de folksonomie a été abordé par de nombreux travaux couvrant une large variété d'approches [2]. Une première catégorie de travaux visent à extraire la sémantique émergente des folksonomies en mesurant la similarité sémantique des tags (Cattuto et al., 2008), ou en exploitant les associations de tags basées via les utilisateurs pour extraire des relations taxonomiques (Mika, 2005). D'autres approches s'appuient davantage sur une contribution des utilisateurs pour *taguer les tags* (Tanasescu et Streibel, 2007), ou pour structurer les tags à l'aide d'une syntaxe simple permettant de spécifier des relations de subsumption (">" ou ">") ou de synonymie ("=")(Huynh-Kim Bang et al., 2008). L'initiative Linked Open Data [3] consiste à proposer un cadre de bonnes pratiques et une série de schémas RDFS (tels que SCOT [4] pour les tags ou SIOC [5] pour les sites sociaux) visant à améliorer l'interopérabilité des plate-formes d'échanges et production de connaissances. Notons également l'ontologie du tagging NiceTag (Limpens et al., 2009) proposant un modèle pivot permettant d'intégrer divers modèles de tags mais aussi rendant compte de la diversité de forme et d'usages des tags, telle que décrit en détails par Monnin et al. (2010). Afin de désambiguïser les tags, Passant et Laublet (2008) proposent MOAT, un modèle permettant de relier les tags avec des URIs de resources du Web décrivant leur signification. D'autres approches intègrent les mesures de similarité dans un processus de mapping entre tags et concepts d'ontologies disponibles en ligne (Specia et Motta, 2007). Enfin, notre approche peut être rapprochée des méthodes de construction d'ontologie

2. Pour un état de l'art complet, voir Limpens (2010, Chapitre 3)
3. linkeddata.org
4. http ://scot-project.org/
5. http ://sioc-project.org/

à partie de textes Aussenac-Gilles et al. (2000), ou de bases de données Golebiowska (2002). Plus récemment,Braun et al. (2007) proposa d'intégrer les processus d'évolution d'ontologies dans l'utilisation quotidienne d'outils de bookmarking par exemple.

L'automatisation complète du processus d'enrichissement de folksonomie reste cependant difficile. Premièrement, les mesures de similarité de (Cattuto et al., 2008; Markines et al., 2009; Specia et Motta, 2007) ou autres méthodes d'extraction de structure sémantique (Mika (2005), Heymann et Garcia-Molina (2006)) sont utiles pour amorcer le processus, mais leur précision est limitée. L'exploitation d'ontologies disponibles en ligne est également limitée par leur nombre restreint et leur pertinence partielle vis à vis des spécificités d'une communauté. D'autre part, les approches s'appuyant exclusivement sur la contribution des utilisateurs risquent d'induire, sans interfaces ergonomiques et adaptées à leurs pratiques, une surcharge cognitive. Par ailleurs, aucune approche, à notre connaissance, ne propose une prise en compte des points de vue divergents quant à la structuration sémantique des tags.

3 Un cycle de vie pour les folksonomies enrichies

Le cycle que nous proposons s'intègre dans l'activité de notre communauté afin de ne pas perturber les tâches courantes des utilisateurs. Notre système consiste à mettre en place une synergie entre automatisations, pour l'amorçage, et participations des utilisateurs, pour réguler et corriger les éventuelles erreurs des automates. La structuration de la folksonomie consiste à spécifier entre les tags des relations sémantiques de thesaurus en suivant le standard SKOS [6] : les variantes orthographiques (`skos:closeMatch`), les relations associatives reliant par exemple "energie" et "electricité" (`skos:related`), et les relations d'hyponymie reflétant le degré relatif de généralité entre deux notions comme par exemple "pollution" qui est une notion *plus générale* (`skos:broader`, elle-même inverse de `skos:narrower`) que "pollution des sols". De plus, notre approche supporte tout au long du cycle les points de vue divergents concernant la structuration des tags. Ceci est permis par le modèle SRTag [7] décrivant les points de vue en associant à chaque relation réifiée l'accord ou le désaccord d'un utilisateur avec cette relation. La réification est effectuées à l'aide de graphes nommés Carroll et al. (2005); Gandon et al. (2007) encapsulant chaque triplet décrivant une relation entre deux tags dans un graphe possédant une URI propre. Ce modèle permet donc de capturer plusieurs relations par paire de tags en gardant la trace des utilisateurs approuvant chacune de ces relations.

Le cycle d'enrichissement commence avec une folksonomie *à plat* (aucun tag relié sémantiquement) et se décompose en 6 étapes :

1. Les traitements automatiques, détaillés en section 4, effectués dans des périodes de basse activité, extraient des relations sémantiques entre les tags.
2. Les utilisateurs individuels (section 5) valident ou corrigent les relations sémantiques calculées ou proposent de nouvelles relations via une interface intégrée dans un outils de navigation. Chaque utilisateur maintient ainsi sa propre structuration de la folksonomie.

6. Simple Knowledge Organisation System http ://www.w3.org/TR/skos-reference
7. http ://ns.inria.fr/srtag/2009/01/09/srtag.html

3. Un autre type d'agent automatique, le solveur de conflits (section 5.2), détecte et propose des solutions temporaires aux incohérences pouvant apparaître entre les points de vue individuels.
4. Les résultats du solveur de conflits sont ensuite mis à profit pour l'utilisateur référent qui est en charge de maintenir une structuration globale et cohérente (section 5.3).
5. A ce stade, nous disposons d'une folksonomie structurée selon plusieurs points de vue. Chaque point de vue peut alors être enrichi avec les contributions des autres utilisateurs à l'aide de règles garantissant une cohérence locale (section 5.4).
6. Un autre cycle redémarre pour prendre en compte les tags nouvellement ajoutés.

4 Extraction automatique de relations sémantiques

Cette section présente la première partie du cycle consistant à extraire les relations entre tags et réalisées par trois algorithmes différents : le premier basé sur des distances syntaxiques, les deux autres sur l'analyse de la structure de graphe triparti de la folksonomie

4.1 Analyse du label des tags (Algorithme. 1)

Cette première méthode consiste à comparer les labels des tags à l'aide d'une combinaison de mesures de similarité morphologique des chaînes de caractères de chaque tag. Ce type de similarité est typiquement employée pour regrouper des variantes orthographiques d'une même notion(Specia et Motta, 2007), mais nous en avons étendu l'usage à la détection d'autres types de relations sémantiques entre tags.

Nous avons comparé les mesures de similarité morphologiques implantées dans le paquet SimMetrics [8] qui donnent pour chaque paire de tag (t_1, t_2) une valeur normalisée entre 0 et 1, 1 signifiant que les deux tags comparés sont identiques. Le but de ce comparatif était d'évaluer la capacité de chaque métrique de SimMetrics à détecter chaque type de relation sémantique. La métrique MongeElkan faisant partie de ce paquet est une métrique hybride qui décompose chaque tag en ses sous-chaînes de caractères, et compare chaque sous-chaînes avec toutes les autres à l'aide d'une métrique tierce. Pour notre expérience, nous avons utilisé 15 métriques et la combinaison des ces 15 métriques avec la métrique de MongeElkan, soit un total de 30 métriques. Notre expérience se base sur un jeu de référence, réalisé manuellement et validé par un expert de l'Ademe, constitué d'une liste de paire de tags reliés par une relation sémantique. Ce jeu de test est divisé en 4 sous-ensembles de 22 paires correspondant chacun à un type de relation (*variante orthographique*, *hyponimie, associative)* plus un jeu de tags non lié sémantiquement. Ce test comparatif se rapporte donc à un problème de recherche d'information où pour chaque métrique, chaque type de relation, et un seuil fixé nous pouvons compter le nombre de vrais positifs (le nombre de paires récupérées et effectivement liées dans le jeu de référence) et le nombre de faux positifs (le nombre de paires récupérés à tort). Ainsi, le comparatif se base sur la précision, le rappel et la moyenne harmonique pondérée F_1 pour chaque métrique et pour chaque type de relation.

Les résultats de ce comparatif nous ont permis de mettre au point une heuristique qui extrait différent types de relations sémantiques entre tags. Cet algorithme proçède à 3 tests pour

8. http ://www.dcs.shef.ac.uk/~sam/stringmetrics.html

chaque paire de tags comparés (t_1, t_2). (1) La métrique MongeElkan-Soundex MES permet de vérifier que la paire de tags comparés est relié par une des 3 relations si $MES(t_1, t_2) > \tau_a$. (2) Ensuite la métrique de JaroWinkler JW détecte si les tags comparés sont des variantes orthographiques lorsque $JW(t_1, t_2) > \tau_b$ (par exemple "transport" et "transports"). (3) Dans la cas contraire, nous utilisons la métrique de MongeElkan-NeedlemanWunch $MENW$ et exploitons son assymétricité en calculant $\delta = MENW(t_1, t_2) - MENW(t_2, t_1)$. Si $\delta \leq -\tau_c$, alors nous en concluons que le tag t_1 est plus particulier que le tag t_2, et si $\delta \geq \tau_c$ alors nous en concluons l'inverse (le tag t_1 est plus général que le tag t_2, comme par exemple "transport" *est plus général que* "transport marchandise"). Les valeurs pour les trois seuils sont déterminées experimentalement. Lorsque deux tags ne sont ni des *variantes orthographiques,* ni reliés par une relation d'*hyponimie,* alors nous en déduisons qu'ils partagent une relation *associative* (comme "transport" et "transfert"*).*

4.2 Analyse de la structure de folksonomie

4.2.1 Algorithme 2

Cattuto et al. (2008) ont montré que les tags ayant des patterns de cooccurrence similaires, c'est à dire les tags n'étant pas nécessairement cooccurrents entre eux mais cooccurrents avec les mêmes tags, tendent à partager des relations de type *associative*. La première étape du calcul de cette similarité consiste à aggréger les données de tagging dans une représentation vectorielle v_i de chaque tag t_i dont les composantes sont données par le nombre de cooccurrences entre le tag t_i et chacun des autres tags t_j lorsque $t_i \neq t_j$ et 0 lorsque $t_i = t_j$. En mettant à 0 la composante du vecteur correspondant à la cooccurrence d'un tag avec lui-même, ceci permet précisément de ne pas prendre en compte la fréquence de coocurrence simple entre les tags. La valeur de la similarité σ entre les tags t_i et t_j est ensuite donnée par $\sigma = \cos(\vec{v_i}, \vec{v_j}) = \frac{\vec{v_i}.\vec{v_j}}{\|\vec{v_i}\|_2.\|\vec{v_j}\|_2}$. Lorsque cette valeur est au dessus d'un seuil déterminé experimentalement, les tags t_i et t_j sont reliés par une relation *associative* (cf exemples en figure 1).

4.2.2 Algorithme 3

Le troisième algorithme exploite l'approche de Mika (2005) visant à exploiter les inclusions d'ensembles d'utilisateurs de certains tags pour en déduire des relations d'*hyponimie*. Soit S_i l'ensemble des utilisateurs du tag t_i, et S_j l'ensemble des utilisateurs du tag t_j. Si l'ensemble S_i est inclus dans l'ensemble S_J, de telle sorte que nous avons $S_i \subset S_j$, avec $card(S_i) > 1$ et $card(S_j) > card(S_i)$, alors nous pouvons inférer que le tag t_j (*e.g.,* "énergie") est *plus général* que le tag t_i (*e.g.,* "électricité").

4.3 Application sur un jeu de données réelles

Nous avons appliqué les trois méthodes présentées dans cette section sur un jeu de données réelles composé des tags des membres de delicious.com ayant utilisé au moins une fois le tag "ademe" (sous-folksonomie *delicious*), des tags extraits des données internes à l'Ademe concernant les thèses financées (sous-folksonomie *thesenet*), et la base d'indexation du corpus Ademe (sous-folksonomie *caddic*). Au total, ce jeu de données contient 9037 tags, 6386

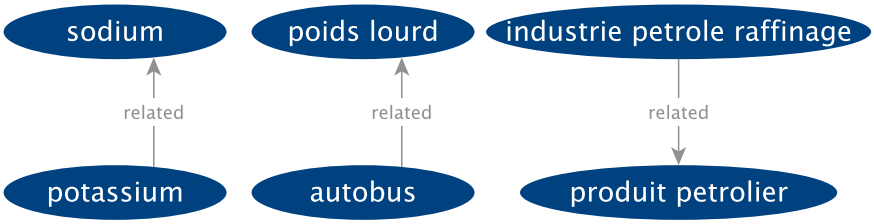

FIG. 1 – *Exemple de relations entre tags calculées par l'algorithme 2.*

ressources, 2238 utilisateurs, et 38690 associations de *tagging, i.e.* "1 tag - 1 utilisateur - 1 ressource".

En terme de nombre de relations calculées, l'algorithme 1 fournit plus de relations (71034) que l'algorithme 2 (8377, dont 97% dans le jeu *delicious*) ou l'algorithme 3 (302 réparties entre *delicious* et *thesenet*). Ceci s'explique par le fait que l'algorithme 1 permet d'établir potentiellement des liens entre tous les tags du jeu de données, en particulier entre les différentes sous-folksonomies, alors que les autres méthodes se basent sur la structure triparti de chaque folksonomie. La répartition des résultats pour l'algorithme 2 s'explique par le fait que, dans *delicious,* les utilisateurs réutilisent plus souvent les tags déjà existants pour taguer un nombre plus restreints de ressources distinctes. En conséquence, les tags ont plus de chances de partager des patterns de cooccurrence similaires avec d'autres tags dans *delicious,* et ainsi plus de chances d'avoir des hautes valeurs de similarité selon l'algorithme 2. Le faible nombre de résultats de l'algorithme 3 s'explique par le fait que les cas d'inclusions de communautés d'utilisateurs de certains tags sont finalement assez rares. Ce point pourrait cependant être amélioré en incluant un taux d'inclusion des ensembles. Le nombre total de relations calculées sur des paires de tags distinctes est de 83027 pour un temps de calcul total de 25647s (dont 20952 pour l'algorithme 1) avec une machine équipé d'un quadri-coeur Core2Duo de 3.00 GHz et 8Go de RAM. Nous précisons également que l'algorithme 1 est le seul incrémental car indépendent de la structure triparti des folksonomies que l'ajout de nouveau tags modifie. Le temps de calcul des itérations suivantes est donc réduit. Afin d'illustrer les résultats donnés par cette étape du cycle, nous donnons en figure 1 un exemple des relations obtenues avec l'algorithme 2.

5 Capture et exploitations des contributions individuelles

5.1 Une interface pour contribuer à la structuration des tags

Grâce au model SRTag nous sommes capables de décrire la structuration de la folksonomie tout en supportant, pour une même paire de tags, de multiples relations associées à différents utilisateurs via des liens d'approbation ou de rejet . Ces points de vue individuels sont capturés grâce à une interface qui intègre des fonctionnalités d'édition des liens sémantiques entre tags suggérés au moment de la recherche d'information. L'idée est en effet de bénéficier de l'activité de recherche des utilisateurs pour les inciter à corriger les relations sémantiques calculées ou proposées par d'autres utilisateurs afin qu'ils puissent maintenir leur propre vision de la structuration de la folksonomie.

FIG. 2 – *Copie d'écran de l'interface intégrant des fonctionnalités d'édition de relations sémantiques (partie gauche) dans un outil de navigation dans la folksonomie (partie droite où les ressources associées au tag cherché, ici "pollution", sont affichées).*

L'interface de navigation dans la folksonomie est présentée figure 2. Dans la partie gauche est située la barre de recherche et les tags suggérés sont répartis dans 4 zones différentes selon le type de relation qu'ils partagent avec le tag cherché. Les ressources associés au tag cherché ainsi qu'aux tags équivalents (*close match*) sont affichés dans la partie droite. L'utilisateur peut ensuite rejeter une relation entre le tag cherché et un tag suggéré en cliquant sur la croix située à côté de chaque tag suggéré. Il est également possible de modifier une relation par une action de cliquer-déposer en déplaçant un tag d'une zone à une des 3 autres zones. Ces manipulations restent donc optionnelles et l'utilisateur peut les ignorer et utiliser uniquement les fonctionnalités de recherche. Chaque utilisateur peut ainsi structurer les relations sémantiques entre tags selon son propre point de vue. Cependant, des incohérences entre points de vue peuvent apparaitre et sont détectées et traitées par un autre module que nous décrivons ci-après.

5.2 Détection et résolution de conflits

5.2.1 Principe du solveur de conflits

Un autre type d'agent automatique, opérant également pendant les périodes de faible activité, recherche et propose des solutions aux conflits émergeant entre les points de vue individuels. Un conflit apparaît lorsque plusieurs différentes relations sont proposées pour une même paire de tags (si un utilisateur change d'avis, nous mettons simplement à jour son point de vue). Par exemple, le tag "pollution" est *plus spécifique* que le tag "co2" pour un nombre n_1 d'utilisateurs, alors que pour un nombre n_2 d'utilisateurs, le tag "pollution" est simplement *associé* (au sens *related* de SKOS) au tag "co2". De plus, un autre nombre n_3 d'utilisateurs peuvent aussi approuver le fait que "pollution" est *plus général* que "co2". Le solveur de conflit compte tout d'abord, pour chaque paire de tags ayant plusieurs relations conflictuelles, le nombre d'approbations $nbApp_i$ pour chaque relation conflictuelle, $i \epsilon [1, n]$ n étant le nombre de relations conflictuelles pour cette paire de tags. Ensuite, il trouve la relation pour laquelle ce nombre est maximum, calcule ce maximum $max\{nbApp_i\}_{i\epsilon[1,n]} = nbApp_{max}$, et compare le ratio $r = \frac{nbApp_{max}}{\sum_n nbApp_i}$ avec une valeur de seuil donnée τ_{cs}. Si ce ratio est supérieur à τ_{cs}, alors le

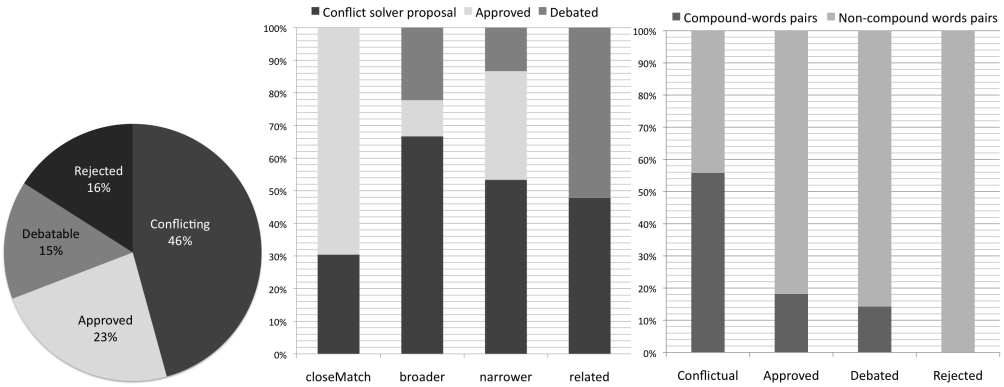

FIG. 3 – *(gauche) Distribution globale des différents cas de conflits. (milieu) Distribution des différents cas de conflit en fonction du type de relation. (droite) Distribution des différents types de paires de tags (avec ou sans noms composés) en fonction des différents types de conflit.*

solveur de conflits approuve la relation correspondante, si, au contraire, ce ratio est inférieur au seuil fixé, alors le solveur de conflits considère qu'aucun consensus n'est atteint et il approuve la relation *associative* (*related* de SKOS) comme compromis entre les points de vue divergents. En effet, cette relation est la plus souple des relations de thesaurus et représente un engagement ontologique moins fort.

5.2.2 Expérience

Afin d'évaluer la nature et le nombre de conflits pouvant apparaître dans un processus de stucturation d'une folksonomie, nous avons conduit une expérience au sein de 5 utilisateurs de l'Ademe à qui nous avons demandé de choisir la relation sémantique la plus pertinente pour une liste de 94 paires de tags (t_1, t_2) parmis les possibilités suivantes : t_1 est une variante orthographique de t_2 (*closeMatch*), t_1 est plus général que t_2 (*broader*), t_1 est plus spécifique que t_2 (*narrower*), t_1 est un terme associé à t_2 (*related*), ou t_1 n'est pas relié à t_2. Ces choix ont ensuite été traduits à l'aide de notre modèle SRTag permettant de décrire une structuration multi-points de vue des tags. Lorsque qu'un utilisateur spécifie que deux tags ne sont pas reliés, ce choix a été traduit en un rejet des autres relations pouvant porter sur ces deux tags. Nous avons ensuite appliqué le solveur de conflits sur ce jeu de test. 4 cas peuvent alors être envisagés concernant la relation entre deux tags :

1. *Approuvée* : lorsque qu'une seule relation a été approuvé par tous les utilisateurs,
2. *Conflictuelle* : lorsque plusieurs relations ont été approuvées pour une même paire de tags,
3. *Débattue* : lorsque une seule relation existe entre deux tags, mais a été à la fois approuvée par certains utilisateurs et rejetée par d'autres,
4. *Rejetée* : lorsqu'une relation a été uniquement rejetée par tous les utilisateurs.

La figure3 montre le détail des résultats de cette expérience.

Le premier graphique (gauche) montre la distribution globale pour les 94 paires de tags du jeu de test des différents cas énumérés ci-dessus. Nous voyons que les cas de conflits où plusieurs relations ont été approuvées pour une même paire de tags représentent 46% des cas, ce qui montre l'intérêt d'une approche multi-points de vue, même au sein d'un collectif restreint.

Le second graphique (milieu) montre la distribution des cas de conflits ou non en fonction du type de relation sémantique concernée pour les 94 paires de tags. Pour les paires de tags où plusieurs relations ont été approuvées (cas *conflictuell*e), nous avons considéré uniquement la relation approuvée par le solveur de conflits, *i.e.* la relation approuvée par une majorité d'utilisateurs ou proposée comme compromis par le solveur de conflits. Dans ce cas nous observons que 70% des cas où la relation *closeMatch* a été retenue, celle-ci a été uniquement approuvée. Ceci montre que cette relation suscite un consensus dans une large majorité des cas. En revanche, dans les cas où la relation *broader* ou la relation *narrower* ont été retenues, celles-ci sont dans une large majorité des cas également soit débattues (cas *débattue*) ou en conflit avec d'autres relations (cas *conflictuelle*). La même observation est également valable pour la relation *related.* Ceci montre donc que ces trois derniers types de relations sont difficilement source de consensus.

Dans le troisième graphique, nous avons examiné l'influence que peut avoir la forme des tags. En particulier nous avons analysé les cas où une paire de tags implique d'un côté un mot, et de l'autre côté un mot composé à partir de ce premier mot (comme par exemple dans "pollution" et "pollution des sols") ou à partir d'une variante (comme dans "pollution" et "détection de polluants"). Ceci concerne 30 des 94 paires de tags. Dans ce graphique nous avons affiché la distribution de ces deux types de paires de tags (*paires avec mots composés* et *paires sans mot*s *composés)* pour les différents types de conflits. Les résultats montrent que les paires ayant des relations conflictuelles sont des paires avec mots composés dans la majorité des cas (56%), signe que la relation sémantique pour les paires avec mots composés est plus souvent sujette à conflits.

5.3 Création d'un point de vue consensuel

Les résultats du solveur de conflits sont ensuite exploités pour aider un utilisateur référent à maintenir un point de vue global et consensuel à partir des contributions individuelles. La première tâche du référent à cet égard consiste à choisir une relation dans les cas de conflits détectés. Ensuite, les autres cas résultant du solveur de conflits peuvent être mis en avant pour aider le référent, comme par exemple les cas de paires de tags ayant fait l'objet d'un consensus soit sur l'approbation unanime d'une relation (cas *approuvée*, cf section 5.2) soit sur le rejet unamine (cas *rejetée).* Le référent est modélisé par une classe spécifique dans le modèle SR-Tag ce qui permet ensuite de reconnaître les relations approuvées par le référent et qui seront ensuite ignorées par le solveur de conflits. Le référent peut donc maintenir son propre point de vue quand à la structuration sémantique de la folksonomie, à la manière des utilisateurs courants, tout en disposant néanmoins d'une interface spécifique lui donnant une vue globale.

5.4 Combinaisons des points de vue individuels

A ce stade du processus d'enrichissement de la folksonomie, nous disposons d'une structuration des tags où plusieurs points de vue co-existent de manière indépendante et où un point de

vue global et consensuel se nourissant des différents points de vue émerge. Une fois le point de vue consensuel établi, il est maintenant possible d'enrichir chaque point de vue individuel avec les relations issues des calculs et des autres utilisateurs en garantissant une cohérence locale de chaque point de vue. Nous détaillons ici la stratégie utilisée pour atteindre cet objectif.

En gardant la trace des différents types d'agents intervenants dans le processus (agents automatiques, utilisateurs individuels, solveur de conflits, et utilisateur référent) nous sommes en mesures d'attribuer un ordre de priorité aux relations destinées à enrichir le point de vue de chaque utilisateur u. La stratégie adoptée consiste à inclure, pour un tag t donné, les relations approuvées par les autres agents en appliquant l'ordre de priorité suivant :

1. l'ensemble R_u des relations approuvées par l'utilisateur *u*.
2. l'ensemble R_{ru} des relations approuvées par l'utilisateur référent, sauf si elles sont en conflits avec l'une des relations de l'ensembleR_u
3. l'ensemble R_{cs} des relations approuvées par le solveur de conflits, sauf si elles sont en conflits avec l'une des relations de l'ensembleR_u, ou R_{ru}
4. l'ensemble R_{ou} des relations approuvées par les autres utilisateurs individuels, sauf si elles sont en conflits avec l'une des relations de l'ensembleR_u, R_{ru}, R_{cs}
5. l'ensemble R_{tc} des relations approuvées par l'agent automatique calculant la sémantique émergente, sauf si elles sont en conflits avec l'une des relations de l'ensembleR_u, R_{ru}, R_{cs}, ou R_{ou}

Cette stratégie permet, lors de la suggestion de tags reliés sémantiquement au tag cherché, d'enrichir le nombre de tags suggérés tout en préservant la cohérence du point de vue de chaque utilisateur. Par exemple, si l'utilisateur courant fait une recherche sur le tag "energie" et qu'il n'a proposé aucune relation portant sur ce tag, il pourra néanmoins bénéficier des relations proposées par les autres utilisateurs. Le point de vue de l'utilisateur référent, ou du solveur de conflits si le référent n'a pas encore traité ce cas de conflits, est utilisé pour arbitrer les cas de relations conflictuelles.

6 Conclusion et discussion

Dans cet article nous avons présenté notre approche de l'enrichissement sémantique de folksonomie qui consiste à structurer les tags à l'aide de relations de thesaurus. Nous proposons un système qui aide, via des suggestions automatiques, les utilisateurs à maintenir leur propre structuration de la folksonomie tout en bénéficiant des contributions des autres utilisateurs. Le cycle de vie de la folksonomie enrichie prend de plus en compte l'activité et les pratiques de nos communautés cibles dont l'Ademe, avec son réseau d'experts produisant des resources annotées et administrées par les documentalistes, est un exemple.

Le cycle démarre par des traitements automatiques permettant d'extraire la sémantique émergente de la folksonomie. Nous avons mis au point une méthode combinant des métriques de similarité morphologique des labels des tags après avoir évalué la capacité de ce type de métrique à distinguer, non seulement les variations orthographiques, mais également d'autres relations sémantiques entre tags comme les relations associatives ou d'hyponimie. Nous avons également adapté d'autres méthodes automatiques basées sur l'analyse de la structure de graphe triparti des folksonomies. Les résultats que nous avons obtenus sur un jeu de données réelles

de l'Ademe montrent la complémentarité entre ces 3 méthodes et la pertinence des relations automatiquement extraites. Le modèle SRTag ainsi que l'interface développée permettent ensuite la coexistence et la combinaison des points de vue divergents concernant les relations entre tags. Ces relations permettent en retour de suggérer des tags semantiquement reliés au tag cherché lors de la navigation et d'enrichir ainsi les resultats de recherche.

Notre approche est implantée dans le système de veille collaborative développé dans le cadre du projet ISICIL [9] et en cours de test au sein de l'Ademe. Dans ce contexte, les données générées lors du cycle de vie de la folksonomie sont utilisées pour la détection de communauté grâce au repérage des utilisateurs partageant des structuration de tags similaires (Ereteo et al., 2009). L'inclusion de la dimension temporelle nous semble également nécessaire afin de mieux comprendre l'évolution et les tendances se dégageant de cette analyse de l'activité des communautés en ligne.

Remerciements

Ce travail a été financé par le projet ISICIL ANR- 08-CORD-011-05.

Références

Aussenac-Gilles, N., B. Biébow, et S. Szulman (2000). Corpus analysis for conceptual modelling. In *Workshop on Ontologies and Texts at Knowledge Acquisition, Modeling and Management, 12th International Conference, EKAW 2000*.

Braun, S., A. Schmidt, A. Walter, G. Nagypál, et V. Zacharias (2007). Ontology maturing : a collaborative web 2.0 approach to ontology engineering. In *CKC*, Volume 273 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Carroll, J. J., C. Bizer, P. Hayes, et P. Stickler (2005). Named graphs, provenance and trust. In *WWW '05 : Proceedings of the 14th international conference on World Wide Web*, New York, NY, USA, pp. 613–622. ACM.

Cattuto, C., D. Benz, A. Hotho, et G. Stumme (2008). Semantic grounding of tag relatedness in social bookmarking systems. In *ISWC '08 : Proceedings of the 7th International Conference on The Semantic Web*, Berlin, Heidelberg, pp. 615–631. Springer-Verlag.

Ereteo, G., M. Buffa, F. Gandon, , et O. Corby (2009). Analysis of a real online social network using semantic web frameworks. In *Proc. International Semantic Web Conference, ISWC'09, Washington, USA*.

Gandon, F., V. Bottolier, O. Corby, et P. Durville (2007). Rdf/xml source declaration, w3c member submission. http ://www.w3.org/Submission/rdfsource/.

Golebiowska, J. (2002). *Exploitation des ontologies pour la memoire d'un projet-vehicule - Methode et outil SAMOVAR.* Ph. D. thesis, Universite de Nice-Sophia Antipolis.

Heymann, P. et H. Garcia-Molina (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical report, Stanford InfoLab.
Anglais

9. http ://isicil.inria.fr

Huynh-Kim Bang, B., E. Dané, et M. Grandbastien (2008). Merging semantic and participative approaches for organising teachers' documents. In *Proceedings of ED-Media 08 ED-MEDIA 08 - World Conference on Educational Multimedia, Hypermedia & Telecommunications*, Vienna France, pp. p. 4959–4966.

Limpens, F. (2010). *Multi-points of view enrichment of folksonomies*. Ph. D. thesis, Université Nice - Sophia Antipolis.

Limpens, F., A. Monnin, D. Laniado, et F. Gandon (2009). Nicetag ontology : tags as named graphs. In *International Workshop in Social Networks Interoperability, Asian Semantic Web Conference 2009*.

Markines, B., C. Cattuto, F. Menczer, D. Benz, A. Hotho, et G. Stumme (2009). Evaluating similarity measures for emergent semantics of social tagging. In *18th International World Wide Web Conference*, pp. 641–641.

Mika, P. (2005). Ontologies are Us : a Unified Model of Social Networks and Semantics. In *ISWC*, Volume 3729 of *LNCS*, pp. 522–536. Springer.

Monnin, A., F. Limpens, F. Gandon, et D. Laniado (2010). Speech acts meet tagging : Nicetag ontology. In *I-SEMANTICS '10 : Proceedings of the 6th International Conference on Semantic Systems*, New York, NY, USA, pp. 1–10. ACM.

Passant, A. et P. Laublet (2008). Meaning of a tag : A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*.

Specia, L. et E. Motta (2007). Integrating folksonomies with the semantic web. In *Proc. of the European Semantic Web Conference (ESWC2007)*, Volume 4519 of *LNCS*, Berlin Heidelberg, Germany, pp. 624–639. Springer-Verlag.

Tanasescu, V. et O. Streibel (2007). Extreme tagging : Emergent semantics through the tagging of tags. In P. Haase, A. Hotho, L. Chen, E. Ong, et P. C. Mauroux (Eds.), *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC2007, Busan, South Korea*.

Summary

Tags freely provided by users of social tagging services are not explicitly semantically linked, and this significantly hinders the possibilities for browsing and exploring these data. On the other hand, folksonomies provide great opportunities to bootstrap the construction of thesauri. We propose an approach to semantic enrichment of folksonomies that integrates both automatic processing and user input, while formally supporting multiple points of view. We take into account the social structure of our target communities to integrate the folksonomy enrichment process into everyday tasks. Our system allows individual users to navigate more efficiently within folksonomies, and also to maintain their own structure of tags while benefiting from others contributions. Our approach brings also solutions to the bottleneck problem of knowledge acquisition by helping communities to build thesauri by integrating the manifold contributions of all their members, thus providing for a truly socio-semantic solution to folksonomy enrichment and thesauri construction.