Les motifs séquentiels au service de la structuration des folksonomies

Sandra Bringay, Maguelonne Teisseire, Julien Gomila, Damien Hoffschir, Thibault Vicaire

Les motifs séquentiels au service de la structuration des folksonomies

Sandra Bringay[1,2], Maguelonne Teisseire[3], Julien Gomila[4], Damien Hoffschir[4] et Thibault Vicaire[4]

[1] Dpt. MIAP, Université Montpellier 3,
sandra.bringay@univ-montp3.fr
[2] LIRMM, CNRS, Montpellier 2
[3] UMR TETIS, Maison de la Télédetection,
maguelonne.teisseire@teledetection.fr
[4] Société EKIOO,
{gomila,vicaire,hoffschir}@ekioo.com

Résumé : Avec l'essor du web 2.0, de nombreux systèmes de tagging social se sont développés. Ils permettent aux internautes de partager des ressources et d'y associer des mots-clés. Dans cet article, nous présentons les travaux réalisés dans le cadre d'une collaboration avec l'entreprise EKIOO qui développe un annuaire d'entreprises collaboratif Ekilink. Nous avons travaillé sur la structuration des mots-clés utilisés par les internautes pour décrire les entreprises et qui forment une folkosonomie, afin d'améliorer l'indexation et la recherche d'informations. Nous avons appliqué les méthodes décrites dans la littérature pour rapprocher les mots-clés, en découvrir de nouveaux et expliciter les relations identifiées entre ces mots-clés. En particulier, nous avons adapté une méthode dédiée à l'enrichissement des ontologies et basée sur la recherche de motifs séquentiels pour la découverte de nouveaux liens labellisés. Les expérimentations sur des jeux de données réelles soulignent la pertinence de notre proposition et ouvrent de nombreuses perspectives.

Mots-clés : Folksonomie, Motifs séquentiels, Web social et collaboratif, Annuaire d'entreprises.

1 Introduction

Ces dernières années, les pratiques de tagging social se sont développées au sein du web social et collaboratif 2.0. Elles rendant les internautes plus actifs au sein des réseaux participatifs. Des sites comme Atpic et Flickr pour le partage de photos, Del.icio.us pour le partage de signets, Wikipedia[1] comme encyclopédie participative, permettent aux internautes de partager des ressources et de les indexer avec des mots-clés. Ces ensembles de mots-clés utilisés par les internautes pour catégoriser des ressources de manière non supervisée sont appelés folksonomies (Limpens et al., 2008), (Passant, 2007). Leur succès vient : (1) du faible coût cognitif lié à leur création. Les internautes n'ont besoin ni de connaissance préexistante, ni de

[1] *atpic.com, www.flickr.com, delicious.com, r.wikipedia.or*

compétence spécifique, ni d'accord entre eux ; (2) de l'autorégulation collective qui rend les folksonomies stables dans le temps. Ce sont les mêmes termes populaires qui sont utilisés sur de longues périodes (Halpin et al., 2007).

L'entreprise EKIOO développe un annuaire d'entreprise collaboratif, Ekilink[2], permettant aux internautes de rechercher et de compléter des informations sur les sociétés et de les recommander dans un esprit communautaire et participatif. EKIOO cherche à améliorer les fonctionnalités d'indexation et de recherche d'informations de ce système et en particulier, la fonctionnalité de recherche d'une entreprise par mots-clés. Lorsque l'utilisateur n'est pas satisfait du résultat d'une requête, l'idée est ici assez classique : il s'agit de lui suggérer, en utilisant une structure de type folksonomie, un nouveau critère plus spécialisé en cas de réponses trop nombreuses ou bien plus général ou proche dans le cas d'un nombre de réponses limité.

Dans ce contexte, nous avons travaillé sur les textes descriptifs des entreprises (textes d'une dizaine de lignes et sites web de l'entreprise si existant) et sur les mots-clés renseignés par l'entreprise et qui forment la folkosonomie. Certains mots-clés sont ambigus avec des problèmes de polysémie (un terme a plusieurs significations, e.g. le terme *alimentation* peut désigner l'activité consistant à se nourrir ou bien un système capable de fournir de l'électricité à un ensemble d'appareils), de variabilité d'écriture (e.g. *alimentation, alimentations*), de synonymie (plusieurs termes sont utilisés pour un même concept, e.g. *alimentation, ravitaillement*). De plus, sans structuration explicite et formelle des relations sémantiques entre les mots-clés, si l'on cherche par exemple des ressources associées au mot-clé *activité*, on ne trouve pas les ressources associées à des concepts plus spécifiques tels que *alimentation* ou *travaux*.

Dans la version initiale du système Ekilink, les mots-clés sont structurés manuellement. Dans le cadre de cette collaboration, nous souhaitons proposer des nouvelles solutions afin d'automatiser autant que possible cette tâche fastidieuse. L'objectif est double : (1) découvrir de nouveaux mots-clés pertinents pour la fonctionnalité de recherche d'une entreprise et identifier des relations entre ces mots-clés et (2) typer ces relations en identifiant des relations d'hyperonymie, d'hyponymie, de synonymie ou bien en proposant des étiquettes pour décrire ces relations. Ces types seront utilisés pour ordonner les suggestions faites aux internautes afin de compléter, affiner ou généraliser leurs recherches.

Dans cet article, nous présentons notre contribution qui se décline selon deux axes : (1) l'utilisation de méthodes décrites dans la littérature pour rapprocher les mots-clés et expliciter les relations identifiées entre ces mots-clés. Ces méthodes sont présentées dans la section 2 ; (2) la définition et mise en œuvre d'une méthode à base de textes et dédiée à l'enrichissement des ontologies. Cette méthode, présentée dans la section 3, est basée sur la recherche de motifs séquentiels et est adaptée au cas des folksonomies. Elle permet de découvrir de nouveaux mots-clés ainsi que des liens labellisés entre les mots-clés. Enfin, nous avons réalisé des expérimentations sur des jeux de données réelles qui sont présentées dans la section 4. Ces expérimentations soulignent la pertinence de notre proposition et ouvrent de nombreuses perspectives détaillées dans la section 5.

---

[2] http://www.ekilink.com/

2 État de l'art et motivations

Il existe de nombreuses méthodes pour l'enrichissement des ontologies, qui sont basées sur la découverte de nouveaux concepts et sur l'extraction de relations sémantiques entre ces concepts. Ces méthodes peuvent être adaptées pour découvrir de nouveaux mots-clés et les rapprocher au sein d'une folksonomie. (Maedche & Staab, 2001) distinguent différentes approches selon le type d'entrée utilisé comme source de connaissances (textes, dictionnaires, bases de connaissances, données semi-structurées et structurées…). Dans cet article, nous nous intéressons aux méthodes à base de textes.

2.1 Rapprocher les termes

Dans le cadre de l'enrichissement d'ontologies, de nombreux auteurs se sont intéressés à l'extraction de termes candidats à partir de méthodes statistiques. Ces méthodes sélectionnent des termes en fonction de leur distribution ou fréquence (Daille, 1996), (Faatz & Steinmetz, 2002), (Parekh et al., 2004), (Xu et al., 2002), (Neshatian & Hejazi, 2004). Elles permettent d'identifier de nouveaux éléments, mais ne permettent pas de les lier entre eux sans une intervention humaine.
Il existe deux types de méthodes pour identifier des relations : (1) des ***méthodes syntaxiques*** basées sur la fonction grammaticale des mots dans les phrases. L'hypothèse est que les informations syntaxiques révèlent des connaissances sémantiques. La plupart de ces méthodes utilisent des patrons syntaxiques (Yansarber et al., 2009) (Sevenson & Greenwood, 2005). Par exemple, l'outil de construction d'ontologies Terminae intègre un module LINGUAE pour rechercher des patrons lexico-syntaxiques dans un corpus (Szulman et al., 2002). L'outil LEXICLASS (Assadi, 1998) regroupe les éléments ayant des positions syntaxiques analogues. L'outil ZELLIG (Habert et al., 1996) regroupe les unités lexicales simples en fonction de relations de dépendances syntaxiques partagées dans un corpus. L'outil ASIUM (Faure & Nedellec, 1998) génère des classes de mots apparaissant dans des contextes similaires ; (2) des ***méthodes statistiques ou d'apprentissage***. On trouve en particulier des méthodes de fouille de données basées sur les règles d'association (Srikant & Agrawal, 1997) qui décrivent des relations entre concepts (Bendaoud, 2006), (Maedche & Staab, 2000), (Stumme et al., 2006). Dans la pratique, la plupart des approches sont mixtes et combinent des méthodes syntaxiques et statistiques. Toutefois, avec ces deux types de méthodes, les auteurs obtiennent des relations entre des concepts, mais ces relations ne sont pas étiquetées.

Il existe des approches similaires dédiées aux folksonomies. Les auteurs rapprochent des termes (1) soit par ce qu'ils correspondent à une ***variation d'écriture*** (e.g. une personne va décrire une entreprise avec le mot-clé *sport* et une autre avec *sports* alors qu'elles ont en tête le même concept). (Specia & Motta, 2007) mesurent pour cela la distance d'édition (Levenshtein, 1966) entre deux mots-clés ; (2) soit par ce qu'ils entretiennent une ***relation sémantique***. Il existe des méthodes basées sur des calculs de cooccurrences et des mesures distributionnelles. (Halpin et al., 2007) utilisent une folksonomie et des calculs de cooccurrences comme point d'entrée à la

construction d'une ontologie. Ils représentent la folksonomie sous la forme d'un graphe. Chaque nœud représente un mot-clé sous forme d'un cercle dont le diamètre dépend de sa fréquence d'apparition. Deux mots-clés sont liés lorsqu'ils sont utilisés ensemble. La longueur des arcs dépend de leur degré de cooccurrence. Ce graphe permet à l'ingénieur de la connaissance de repérer les relations sémantiques entre les concepts. (Cattuto et al., 2008) et (Mika, 2005) utilisent un approche distributionnelle en prenant en compte les utilisateurs pour structurer une folksonomie. À partir d'un modèle en trois parties représentant les ressources, les mots-clés et les utilisateurs, ils associent les mots-clés en rapprochant ceux qui ont le plus d'utilisateurs ou de ressources en commun. Par exemple si de nombreux internautes décrivent une entreprise en utilisant les termes *tennis* et *raquette*, on suppose qu'il existe une relation entre ces termes.

2.2 Expliciter les relations entre les termes

Si les méthodes précédentes sont intéressantes car elles permettent d'identifier automatiquement des relations entre mots-clés (e.g. entre *traiteur* et *pâtisserie*), elles n'explicitent pas les liens établis. (Specia & Motta, 2007), à partir de clusters de mots-clés construits selon leur cooccurrence, recherchent dans des ontologies du Web sémantique, la présence de liens pour tous les couples de mots-clés d'un cluster. Une automatisation de cette méthode a été décrite par (Angeletou et al., 2007). (Cattuto et al., 2008) ont également utilisé cette méthode pour étudier les types de liens présents dans la folksonomie.

Toutefois, ces méthodes n'explicitent pas toutes les relations entre les mots-clés (Limpens et al., 2009). En effet, certains sont mal orthographiés, correspondent à du vocabulaire actuel (non présent dans les sources de connaissances) et sont formulés dans différentes langues (peu de sources sont multilingues). Par ailleurs, les internautes utilisent parfois des instances comme mot-clé. Par exemple, ils annotent une photo avec le nom d'une personne qui n'existe pas dans les sources de connaissances. De même, ils utilisent souvent des termes spécifiques. Or, il existe peu de ressources spécifiques et des ressources termino-ontologiques comme Wordnet, utilisée par (Cattuto et al., 2008), contiennent uniquement des termes généraux. (Limpens et al., 2009) proposent d'utiliser une méthode basée sur l'expertise des utilisateurs en le faisant participer à la structuration de la folksonomie. Ils cherchent à limiter l'effort de contribution nécessaire à la formalisation de cette expertise mais l'implication de l'utilisateur reste très importante.

Nous proposons dans cet article de compléter ces approches avec une adaptation de la méthode décrite par (Di-jorio et al., 2008) pour l'enrichissement d'ontologies. Cette méthode identifie des relations et les nomme sans faire appel à des ressources du domaine. Elle est basée sur l'extraction de motifs séquentiels qui mettent en évidence des schémas fréquents sous la forme de séquences. Appliqués sur des textes, ils permettent de découvrir des séquences de mots fréquemment associés dans un ordre donné. Nous décrivons plus en détail cette méthode dans la section suivante.

3 Les motifs séquentiels au service de la structuration des folksonomies

À notre connaissance, les motifs séquentiels n'ont encore jamais été utilisés pour identifier des relations entre les termes d'une folksonomie. L'originalité de l'approche consiste à proposer des relations déjà labellisées simples à intégrer l'application. Dans cette section, nous décrivons les motifs séquentiels et détaillons les différentes étapes.

3.1 Les motifs séquentiels

Introduit par (Agrawal & Srikant, 1995), les motifs séquentiels sont des corrélations entre événements s'enchaînant selon une relation d'ordre (e.g., le temps). Une séquence S est une liste ordonnée non vide d'itemsets s_j notée $S =< s_1\ s_2 \ldots s_n >$. Un itemset est un ensemble non vide d'items i_j noté $(i_1 , i_2 , \ldots , i_n)$. Si la séquence S est incluse dans S', alors S est une sous-séquence de S'. Une base de séquences DB est un ensemble de paires (id,S) où id est l'identifiant de la séquence S. Une paire (id,S) supporte une séquence Sα si Sα est une sous-séquence de S ($S\alpha \preccurlyeq S$). Le support d'une séquence Sα est défini comme le pourcentage de paires qui supportent Sα. Sα est fréquente si son support est au moins égal à une valeur minimale minSup. La recherche de motifs séquentiels consiste à trouver toutes les séquences de longueur maximale dont le support est supérieur à minSup. Plusieurs algorithmes efficaces ont été proposés (Agrawal & Srikant, 1995), (Masseglia et al., 1998), (Zaki, 2001).

3.2 Description des différentes étapes de la méthode

Notre méthode (*cf.* fig. 1) se décompose en 4 étapes :

Étape 1 - Prétraitement des textes :

1. Normalisation : cette étape permet de retirer toutes les marques de ponctuation, majuscules… ainsi que tous les termes contenus dans une stop liste et correspondant aux mots courants utilisés dans la langue française. Par exemple, la phrase « Le Sushi House est un restaurant japonais situé sur la charmante place Chabaneau » est réduite aux mots : « Sushi House est restaurant japonais situé charmante place Chabaneau ».

2. Lemmatisation : l'analyse lexicale permet de remplacer chaque mot par une entité appelée lemme (forme canonique) suivi de sa catégorie grammaticale[3]. Dans l'exemple précédent, nous obtenons : « Sushi@nom être@ver restaurant@nom japonais@adj situer@ver charmant@adj place@nom Chabaneau@nom ». Certaines catégories grammaticales sont supprimées (e.g. conjonctions, pronoms…).

3. Filtrage : afin de ne garder que les termes les plus représentatifs, nous avons utilisé la mesure *idf* proposée dans (Robertson & Jones, 1988) qui fait ressortir l'importance d'un mot dans un corpus. Finalement, nous obtenons : « Sushi@nom restaurant@nom japonais@adj situer@ver charmant@adj place@nom ».

[3] Nous avons utilisé un étiqueteur grammatical de type treetagger

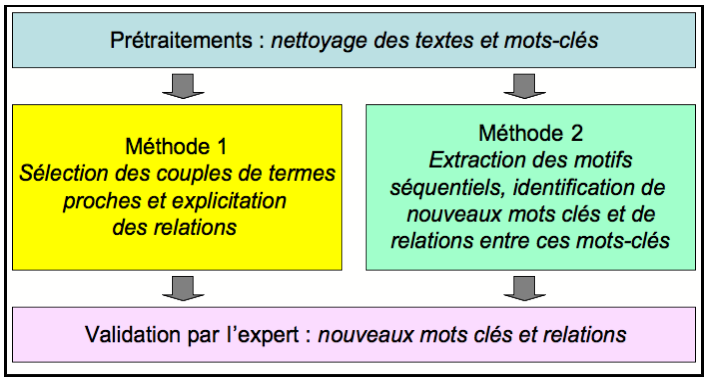

Fig. 1 – Processus de structuration de la folksonomie

Étape 2 – Recherche de liens entre les mots-clés via les mesures distributionnelles et explicitation via des ressources disponibles en ligne

1. Sélection des couples de termes proches : Les entreprises sont décrites par des textes courts et contenant peu/pas de répétitions. Nous ne nous sommes donc pas intéressés aux cooccurrences simples mais uniquement aux méthodes distributionnelles. Soit THEME(Tj) (resp. THEME(Tk)) un thème regroupant tous les textes associés au mot-clé Tj (resp. Tk). Si THEME(Tj) et THEME(Tk) ont beaucoup de ressources en commun, c'est-à-dire si card((THEME(Tj) ∩THEME(Tk)))>NbTexteMin avec NbTexteMin un seuil défini expérimentalement, alors Tj et Tk sont considérés proches. De même, soit CI(Tj) (resp. CI(Tk)) une communauté d'intérêt regroupant tous les internautes utilisant le mot-clé Tj (resp. Tk). Si CI(Tj) et CI(Tk) ont beaucoup d'acteurs en commun, c'est-à-dire si card((CI(Tj)∩CI(Tk)))>NbUtMin avec NbUtMin un seuil défini expérimentalement, alors Tj et Tk sont considérés comme proches.

2. Explicitation des relations : Grâce à l'étape précédente, nous avons créé des couples de termes. Nous les explicitons en utilisant le web service disponible via le portail terminologique TermSciences[4] pour rechercher si deux termes d'un couple sont liés par une relation de parenté. Lorsqu'une relation de parenté existe, c'est-à-dire si les deux termes ont un ancêtre commun assez proche, nous identifions une relation et utilisons le terme du concept commun dans la hiérarchie pour décrire cette relation.

Étape 3 – Recherche de relations étiquetées entre les mots-clés via motifs séquentiels

1. Extraction des motifs séquentiels : Dans notre contexte, la base de séquences est constituée de paires correspondent aux fiches descriptives ou site web des entreprises associées à un identifiant. Les phrases dans ces fiches sont les séquences de données. Comme (Plantevit & Charnois, 2009), nous considérons que les items sont les lemmes associés à leur étiquette morpho-syntaxique et qui sont obtenus à partir des mots dans les phrases des fiches. La relation d'ordre correspond à l'ordre des mots dans la phrase. Par exemple, la séquence < (traiteur@nom) (produire@verbe

[4] http://www.termsciences.fr : ce portail, créé par l'INIST, LORIA et ATILF, mutualise des ressources terminologiques des organismes publics de recherche et d'enseignement supérieur.

viennoiserie@nom) > 10% signifie que dans 10% des textes, on retrouve une phrase contenant le nom *traiteur* suivi d'une deuxième phrase composée du verbe *produire* et du nom *viennoiserie*.

2. Identification de relations entre les mots-clés de la folksonomie : parmi tous les motifs obtenus, on sélectionne ceux qui sont composés d'au moins deux mots-clés déjà présents dans la folksonomie. On utilise alors l'ordre des mots dans la phrase et les étiquettes grammaticales pour choisir l'étiquette qui va représenter la relation entre les deux mots-clés. Pour cela, on préfère tout d'abord les motifs composés de couples de mots-clés situés dans le même itemset (fréquemment associés dans une phrase) à ceux qui ne sont pas dans le même itemset (fréquemment associés mais dans des phrases différentes). Dans les deux cas, on propose au gestionnaire le ou les verbes de l'itemset comme étiquette de la relation puis à défaut les noms et autres lemmes. Par exemple, le motif précédent permet d'identifier une relation entre *viennoiserie* et *traiteur* que l'on étiquette avec le verbe *produire*.

3. Identification et sélection de nouveaux mots-clés : Pour tous les mots-clés de la folksnomie, nous construisons l'ensemble de termes appartenant à leur voisinage en utilisant les relations identifiées via les motifs séquentiels. Pour cela, nous utilisons une mesure de proximité basée sur le support des motifs séquentiels et proche de celle adoptée dans (Di-jorio et al., 2008). Le principe consiste à utiliser une mesure de « ranking » afin de mettre une priorité sur les mots-clés candidats. La formule de calcul associée (formule 1) retourne pour un mot-clé c_0 de la folksonomie et un terme i, la valeur maximum parmi les fréquences d'apparition ordonnée des co-occurrences du mot-clé c_0 et du terme i. Nous choisissons ensuite comme candidats pour le voisinage du mot-clé c_0, les termes i ayant la plus grande proximité.

$$Prox(c_o, i) = \max\left(\begin{array}{l} max\left(\frac{Freq([(i\ c_o)])}{Freq([(c_o)])}, \frac{Freq([(i\ c_o)])}{Freq([(i)])}\right), \\ max\left(\frac{Freq([(i)(c_o)])}{Freq([(i)])}, \frac{Freq([(i)(c_o)])}{Freq([(c_o)])}\right), \\ max\left(\frac{Freq([(c_o)(i)])}{Freq([(c_o)])}, \frac{Freq([(c_o)(i)])}{Freq([(i)])}\right) \end{array}\right) \quad (1)$$

De la même façon et grâce au caractère ordonné des termes dans les motifs séquentiels, nous pouvons mettre en évidence le sens des nouvelles relations. Par exemple, ce n'est pas la *viennoiserie qui produit le traiteur*, mais le *traiteur qui produit la viennoiserie*.

Étape 4 – Validation par le gestionnaire de l'application

L'ensemble des nouvelles relations, obtenues via les mesures distributionnelles et les motifs séquentiels, sont présentées au gestionnaire de l'application qui les valide ou non afin de préserver la cohérence des mots-clés et des relations pré-établies. Cela implique par exemple de ne pas ajouter des relations redondantes, en particulier des relations existantes chez les ancêtres des concepts concernés.

4 Expérimentations

4.1 Description des données et préparation des mots-clés et textes

Avant cette collaboration, le nombre de mots-clés composant la folksonomie et gérés manuellement par l'entreprise EKIOO était de 705. La folksonomie est structurée en 9 parties, correspondant à des domaines d'activité différents des entreprises. De nouveaux mots-clés sont ajoutés manuellement à la folksonomie lorsque EKIOO intègre un nouveau domaine d'activité ou bien lorsqu'une entreprise se décrit avec une nouvelle spécialité. Cet ensemble de mots-clés est évidemment difficile à gérer sur le long terme sans une automatisation du processus de structuration et c'est ce constat qui a motivé la collaboration.

Nous avons travaillé sur deux domaines d'activité particuliers : « alimentation et restauration » et « travaux et habitat ». Le premier domaine était associé à 67 mots de la folksonomie et le second à 91 mots-clés. Nous avons sélectionné des textes issus des fiches descriptives des entreprises et des sites web associés à ces entreprises (248 textes pour le premier domaine et 11818 pour le second).

Nettoyage des mots clés : la mesure de Leveinstein a été appliquée à tous les couples de termes composant la folksonomie afin d'identifier les variations d'écriture. Lorsque la distance était inférieure à 90% (seuil défini empiriquement), nous avons soumis le couple de termes au gestionnaire qui a validé ou non la variation d'écriture. Ce dernier a ainsi pu détecter des fautes d'orthographe, frappe, des mots au pluriel...

Nettoyage des mots textes : A l'issue des étapes de normalisation, de lemmatisation et de filtrage, nous avons obtenu 568 lemmes pour le premier domaine et 1277 pour le second, associés à une étiquette grammaticale. Ces lemmes ont été utilisés en entrée du processus. L'intersection entre les lemmes et les mots-clés de folksonomie initiale est de 47 pour le premier domaine et de 88 pour le second.

4.2 Résultats

Étape 2 : Pour identifier des relations entre des couples de termes et les expliciter, nous avons appliqué l'étape 2 sur la folksonomie initiale et obtenu une centaine de couples associés via les cooccurrences. Nous identifions essentiellement des relations de domaine (comme *parquet* et *flottant*), pertinentes à suggérer à l'utilisateur pour l'aider à préciser sa recherche. Une perspective consiste non pas à considérer comme proche les termes dont l'intersection des thèmes ou des communautés est importante mais ceux dont les thèmes ou communautés sont inclus, afin de faire émerger différents niveaux de généralités entre mots-clés. Nous avons utilisé le portail terminologique TermSciences pour expliciter ces relations grâce à un service qui teste si un terme est plus générique qu'un autre. Nous avons identifié par exemple que le terme *tennis* est plus spécifique que *sport*. Toutefois, nous n'avons réussi à expliciter que peu de couples (28%). En effet, tous les termes identifiés dans les textes n'existent pas dans TermSciences et tous les couples ne sont pas liés par une relation de spécialisation (limitation du web service). Pour améliorer ces résultats, des

ressources autres que TermSciences doivent être envisagées pour prendre en compte plus de termes et d'autres types de relation que celles de spécialisation.

Les résultats des étapes 3.2 et 3.3 sont résumés dans la table 1.

Étape 3.2 : Afin d'identifier des relations entre les termes de la folksonomie, nous avons extrait des motifs séquentiels avec un support minimum très bas (0.016 pour le premier corpus et 0,013 pour le second) afin d'obtenir un maximum de motifs. Nous les avons ensuite filtrés pour conserver ceux contenant deux mots-clés de la folksonomie. Nous avons ensuite présenté des échantillons de ces motifs au gestionnaire qui les a validés. La précision est très bonne : 98% pour le premier corpus et 84% pour le second. Presque tous les motifs peuvent être utilisés pour enrichir la folksonomie, soit par une nouvelle relation, soit par un nouveau mot-clé. Cette étape a notamment permis de rajouter des niveaux structurant dans la folksonomie. Par exemple, on trouve sous le terme « alimentation et restauration », les termes *bar, restaurant, pâtisserie, viande*... Les motifs ont fait ressortir deux sous-ensembles : *bar* avec *restaurant* et *pâtisserie* avec *viande*. Nous avons suggéré au gestionnaire de rajouter deux nouveaux nœuds : *établissement* comme père de *bar* et *restaurant* et *aliment* comme père de *pâtisserie* et *viande*. Les motifs ont également permis de dégager de nouvelles relations autres que des relations de spécialisation comme le lien entre *traiteur* et *viennoiserie* unis par la relation sémantique *produire*.

Étape 3.3 : Nous avons appliqué l'étape 3.3, pour identifier de nouveaux mots-clés et les lier à ceux existant dans la folksonomie. A partir des listes de motifs séquentiels obtenus précédemment, nous avons sélectionné ceux contenant un seul terme de la folksonomie et construit leur voisinage afin de présenter au gestionnaire les 10 termes les plus proches. Nous identifions essentiellement ici des relations de domaine, par exemple l'association entre *pompe* et *chaleur*. La précision de la méthode est ici moins bonne, mais peut être améliorée en triant les motifs proposés au gestionnaire en fonction de l'étiquette grammaticale des mots. Maintenant que la folksonomie a été enrichie, il serait intéressant d'appliquer de nouveau l'étape 3.2 pour identifier de nouvelles relations.

Ces deux étapes devraient être testées sur l'ensemble des motifs ainsi que sur d'autres corpus pour une validation plus complète. Le temps passé par le gestionnaire sur cette évaluation devrait également être estimé.

Table 1. Synthèse des résultats obtenus sur les deux corpus. Les valeurs de précision ont été calculées à partir d'échantillons.

Corpus	**Textes**	**Mots-clés**	**Lemmes**	**lemmes ∩ mots clés**	**Motifs**	**Motifs avec 2 mots-clés**	**Motifs validés**	**Précision %**	**Motifs avec 1 mot-clé**	**Motifs validés**	**Précision %**
Alimenta. Restaurat.	248	67	568	47	6199	1335	1308	98%	407	183	45%
Travaux Habitats	11818	91	1277	88	64496	6449	5417	84%	2056	1089	53%

5 Conclusions et perspectives

Dans cet article, nous sommes intéressés aux folksonomies qui se basent sur l'indexation libre de mots-clés par les internautes et qui offrent de nouveaux supports à la navigation sociale. Nous avons proposé et mis en œuvre une méthode de structuration de folksonomie que nous avons appliquée à celle utilisée par l'entreprise EKIOO pour son système d'annuaire Ekilink qui associe des mots-clés à des entreprises. Nous avons utilisé des méthodes classiques de la littérature, basées sur les cooccurrences de termes et sur l'interrogation de ressources du web sémantique, pour rapprocher des mots-clés, expliciter et nommer ces relations. Nous avons également défini et mis en œuvre une méthode basée sur la recherche de motifs séquentiels pour découvrir de nouveaux mots-clés ainsi que de nouvelles relations entre les mots-clés. Nous avons expérimenté ces méthodes sur des jeux de données réelles et enrichit la folksonomie de l'entreprise EKIOO. Notre approche, en permettant d'identifier des relations de domaine étiquetées et en faisant émerger de nouveaux niveaux structurant dans la folksonomie est complémentaire des approches existantes qui détectent des relations proches du thésaurus comme les méthodes basées sur les distances d'édition (Specia & Motta, 2007) ou sur les approches distributionnelles (Mika, 2005).

Notre approche présente l'avantage d'être indépendante de la structure initiale de la folksonomie et des textes utilisés en entrée du processus. Par ailleurs, nous utilisons des méthodes et des ressources linguistiques simples (étiqueteur grammatical de type TreeTagger pour l'essentiel). Nous avons privilégié une approche de ce type pour les raisons suivantes : (1) Les outils linguistiques plus complexes tels que les analyseurs syntaxiques (cf. campagne EASY[5]) peuvent se révéler peu adaptés pour traiter des textes de domaines de spécialité comme ceux présentés dans cet article. Le traitement, pouvant se révéler long et complexe, est en outre très dépendant des langues. (2) Contrairement à la plupart des approches syntaxiques qui produisent des relations binaires (sujet-verbe, verbe-objet, etc.), nous obtenons des relations ternaires voire plus complexes, utilisées pour suggérer des étiquettes et nommer les relations.

Ainsi, nous avons privilégié une approche basée sur le calcul du support des motifs séquentiels représentant des associations fréquentes de termes dans les textes. D'autres mesures d'intérêt comme la longueur du motif ou la présence d'items particuliers (comme le verbe *être* ou *faire*) pourraient être utilisées. Nous complétons cette approche avec uniquement des connaissances linguistiques de base. En particulier, nous conservons certaines spécificités de la langue comme l'ordre des mots dans les textes. Cette information, présente dans les motifs, nous permet de les ordonner lorsque nous les présentons au gestionnaire de l'application pour validation.

Nous rencontrons les mêmes limites que les approches traditionnelles concernant la synonymie, la polysémie et les variations d'écriture. Toutefois, en repartant des textes pour étiqueter les relations, nous nous émancipons des problèmes de vocabulaires trop récents, trop spécialisés ou multilingues pour être présents dans les thésaurus ou autres ressources comme dans l'approche de (Specie & Motta, 2007). Une limite essentielle de notre approche reste la validation fastidieuse par le

[5] http://www.limsi.fr/Recherche/CORVAL/easy/

gestionnaire. Une perspective intéressante serait d'intégrer une approche semblable à celle de (Béchet 2009) qui propose un protocole d'évaluation visant à noter et à ordonner automatiquement des relations syntaxiques dites induites, en couplant l'interrogation d'un moteur de recherche sur le Web avec diverses mesures statistiques (information mutuelle, coefficient de Dice et fréquence, popularité…).

Finalement, à partir d'une pratique individuelle consistant à associer des mots clés à des ressources, une approche comme celle présentée dans cet article, permet aux internautes de coopérer de manière implicite, grâce à la structuration déduite par le système, en coordonnant a posteriori leur recherche d'information.

Remerciements

Nous remercions Cécile Ochman qui a développé les premières briques du prototype ainsi que Mathieu Roche pour ses remarques et ses conseils avisés.

Références

ANGELETOU S. (2008). Semantic Enrichment of Folksonomy Tagspaces. *International Semantic Web Conference*, p. 889-894.

AGRAWAL R. & SRIKANT R. (1995). Mining Sequential Patterns. *11th IEEE International Conference on Data Engineering*, p. 3–14.

ASSADI H. (1998). Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires. *Thèse de l'Université Paris 6*.

BENDAOUD R. (2006). Construction et enrichissement d'une ontologie à partir d'un corpus de textes. *RJCRI'06*, p. 353–358.

BECHET N. (2009). Description d'un protocole d'évaluation automatique comme alternative à l'évaluation humaine. Application à la validation de relations syntaxiques induites. *Evaluation des Méthodes d'Extraction de Connaissances dans les Données, Atelier de EGC'09.*

CATTUTO C., BENZ D., HOTHO A. & STUMME G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. ISWC2008, LNCS 5318, p. 615-631.

DAILLE B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *P. Resnik and J. Klavans (eds). The Balancing Act : Combining Symbolic and Statistical Approaches to Language, MIT Press*, p. 49–66.

DI-JORIO L., FIOT C., ABROUK L., HERIN D., TEISSEIRE M. (2008). Sequential Patterns for Maintaining Ontologies over Time. OTM Conferences (2) 2008: 1385-1403

FAATZ A. & STEINMETZ R. (2002). Ontology enrichment with texts from the WWW. *The Semantic Web Mining Conference (WS'02)*.

FAURE D. & NEDELLEC C. (1998). ASIUM: Learning subcategorization frames and restrictions of selection. *ECML'98 Workshop on Text Mining*, Chemnitz.

HABERT B., NAULLEAU E. & NAZARENKO A. (1996). Symbolic word clustering for medium-size corpora. *International Conference on Computational Linguistics (COLING'96), p. 490-495.*

LEVENSHTEIN V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.,* 10(8), p. 707–710.

LIMPENS F., GANDON F. & BUFFA M. (2008). Rapprocher les ontologies et les folksonomies : un Etat de l'art. Actes des 19ème journées d'Ingénierie des Connaissances, LORIA, Nancy.

LIMPENS F., GANDON F. & BUFFA M. (2009). Sémantique des folksonomies: structuration collaborative et assistée. Actes des 20ème journées d'Ingénierie des Connaissances, p. 37-48.

HALPIN H., ROBU V. & SHEPHERD H. (2007). The complex dynamics of collaborative tagging. *WWW 2007*, p. 211-220.

MAEDCHE A & STAAB S. (2001). Ontology Learning for the Semantic Web. *IEEE Intelligent Systems, Spe. Issue on the Semantic Web*, 16(2).

MASSEGLIA F., CATHALA F. & PONCELET P. (1998). The PSP approach for mining sequential patterns. *The Second European Conference on Principles of Data Mining and Knowledge Discovery*, p. 176–184.

MIKA P. (2005). Ontologies are Us : a Unified Model of Social Networks and Semantics. *ISWC, volume 3729 of LNCS*, p. 522–536.

NESHATIAN K. & HZJAZI M. R. (2004). Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies. *2nd Workshop on Information Technology and its Disciplines,* p. 43–48.

PAREKH V., GWO J-P. & FININ T. (2004). Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. *International Conference of Information and Knowledge Engineering*.

PASSANT A. (2007). Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs. *International Conference on Weblogs and Social Media*.

PLANTEVIT M. & CHARNOIS T. (2009). Motifs séquentiels pour l'extraction d'information : illustration sur le problème de la détection d'interactions entre gènes. *Traitement Automatique des Langues Naturelles* (TALN), Senlis, France.

ROBERTSON S. E. & JONES K. S. (1988). Relevance weighting of search terms. p.143–160.

SPECIA & MOTTA E. (2007). Integrating folksonomies with the semantic web. *4th European Semantic Web Conference.*

STEVENSON, M. & GREENWOOD M. (2005). A Semantic Approach to IE Pattern Induction. *43rd Meeting of the Association for Computational Linguistics (ACL-05), Ann Arbour, Michigan*, p. 379-386.

SRIKANT R. & AGRAWAL R. (1997). Mining generalized association rules. *Future Generation Computer Systems,* 13(2–3): 161–180, 1997.

SRIKANT R. & AGRAWAL R. (1996). Mining sequential patterns: Generalizations and performance improvements. *5th International Conference on Extending Database Technology* (EDBT '96), p. 3–17.

STUMME G., HOTHO A. & BERENDT B. (2006). Semantic web mining: State of the art and future directions. *Web Semantics : Science, Services and Agents on the World Wide Web*, 4(2), p. 124–143.

SZULMAN S., BIEBOW B. & AUSSENAC-GILLES N (2002). Structuration de terminologies á l'aide d'outils de tal avec TERMINAE. *A. Nazarenko and T. Hamon, editors, Traitement automatique des langues. Structuration de terminologie*, volume 43, p. 103-128.

YANGARBER R., GRISHMAN R. & TAPANAINEN P. (2000). Unsupervised Discovery of Scenario-Level Patterns for Information Extraction. *6th ANLP, Seattle*, p. 282-289.

ZAKI M.J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning,* 42(1/2), p. 31–60.

XU F., KURZ D., PISKORSKI J. & SCHMEIER S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. *3rd international conference on language resources and evaluation*.