REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Djillali Liabes

Faculté de Technologie

Département d'Informatique

Traitement sémantique des données semi-structurées et non structurées : Des folksonomies aux ontologies

THÈSE

présentée et soutenue publiquement

pour l'obtention du

Diplôme de Doctorat en Sciences
(spécialité informatique)
(Option : Systèmes d'Informations et de Connaissances)

par

Zahia MAROUF

Composition du jury

Président :	Ahmed Lehireche, Professeur, Université Djillali Liabès, Sidi-Bel-Abbès
Directeur de Thèse :	Sidi Mohamed BENSLIMANE, Professeur, Unmiversité Djillali Liabès, Sidi-Bel-Abbès
Examinateurs :	Ghalem BELALEM, Professeur, Université d'Oran, Oran Mimoun MALKI, Professeur, Université Djillali Liabès, Sidi-Bel-Abbès Fatima DEBBAT, Maitre de Conférences, Université de Mascara, Mascara Reda Mohamed HAMOU, Maitre de Conférences, Université Taher Moulay, Saida

A l'âme de mon père,

Remerciements

Je tiens à exprimer mes plus vifs remerciements à Sidi Mohamed Benslimane qui fut pour moi un directeur de thèse attentif et disponible malgré ses charges bien nombreuses. Sa compétence, sa rigueur scientifique et sa clairvoyance m'ont beaucoup appris. Ils ont été et resteront les moteurs de mon travail de chercheur.

J'exprime tous mes remerciements à l'ensemble des membres de mon jury :Mademoiselle Debbat Fatima, et Mesieurs Lehireche Ahmed, Malki Mimoune, Belalem Ghalem, et Hamou Reda Mohamed qui ont accepté d'évaluer ce travail.

J'adresse toute ma gratitude à tous mes ami(e)s et à toutes les personnes qui m'ont aidé dans la réalisation de ce travail.

Je tiens à remercier toutes les personnes qui ont bien voulu me rencontrer et répondre à mes questions et sans qui ce travail n'aurait pas pu voir le jour.

Enfin, j'adresse toute mon affection à ma famille, son intelligence, sa confiance, sa tendresse, son amour me portent et me guident tous les jours.

Enfin, je remercie Khaled, avec qui je partage ma vie, qui m'a supportée pendant la finalisation de ce travail et a su m'encourager dans les moments les plus difficiles.

Une pensée pour terminer ces remerciements pour toi qui n'a pas vu l'aboutissement de mon travail mais je sais que tu en aurais été très fier.

Résumé

Le Web 2.0 est l'évolution du Web vers plus de simplicité et d'interactivité où l'utilisateur est au centre de service en termes de publications et de réactions. Cela fait passer l'utilisateur du statut de consommateur à celui de producteur. Les folksonomies constituent des fonctionnalités phares du Web 2.0. Elles permettent aux utilisateurs de décrire des ressources sur le web (billet de blog, page Web, photos, vidéos...) par des mots-clés choisis librement. Bien que les folksonomies et leurs tags soient subjectifs et dépendant du contexte, ce qui rend difficile leur exploitation, de nombreux chercheurs ont prouvé l'existence d'une sémantique implicite dans ces données non structurées. cette thèse propose une approche pour extraire des structures ontologiques à partir des folksonomies. L'approche exploite la puissance du clustering flou, et emploie de nouvelles mesures de similarité et de généralité.

Le processus de clustering flou détecte les tags ambigüs et les désambiguïse à la fois, la nouvelle mesure de similarité est utilisée pour découvrir les relations entre les tags, et la mesure de généralité est employée pour extraire la structure hiérarchique de la folksonomie. Les nouvelles mesures donnent des résultats plus précis car elles calculent les co-occurrences tout en prenant en compte les utilisateurs. L'ontologie générée peut être utilisée pour améliorer diverses tâches, telles que l'évolution et l'enrichissement des ontologies par l'ajout de nouveaux concepts créés par les communautés d'internautes. Elle peut également être utilisée dans la recommandation des tags pour guider le processus de marquage en proposant des tags plus contrôlées. En outre, l'ontologie générée peut être employée dans l'expansion des requêtes pour améliorer les la recherche de l'information.

Mots-clés: folksonomies, ontologies, clustering flou, désambiguïsation, mesure de similarité, mesure de généralité.

Abstract

Web 2.0 is an evolution toward a more social, interactive and collaborative web, where user is at the center of service in terms of publications and reactions. This transforms the user from his old status as a consumer to a new one as a producer. Folksonomies are one of the technologies of Web 2.0 that permit users to annotate resources on the Web. This is done by allowing users to use any keyword or tag that they find relevant. Although folksonomies require a context-independent and inter-subjective definition of meaning, many researchers have proven the existence of an implicit semantics in these unstructured data. This thesis proposes an approach for extracting ontological structures from folksonomies that exploits the power of fuzzy clustering using new similarity and generality measure.

The fuzzy clustering process discovers ambiguous tags and disambiguates them all at once, the new similarity measure is used to extract relations between tags, and the generality degree is employed to extract the hierarchical structure from the folksonomy. The new measures give more accurate results as they calculate co-occurrences based on distinct users and not only in the number of times when two words co-occur. The generated ontology can be used to enhance various tasks such as ontology evolution and enrichment by adding new concepts created by the community to the ontology. It can be also used in tag recommendation to guide the tagging process by proposing more controlled tags. Moreover, the generated ontology can be employed in query expansion to improve retrieval performance in information retrieval operations.

Keywords: folksonomies, ontologies, fuzzy clustering, disambiguation, similarity measure, generality measure.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Dans ce chapitre nous explicitons le sujet traité comme point de départ. Ensuite, nous posons la problématique et présentons nos contributions. Enfin nous donnons le plan de la thèse

1.1 Contexte de la thèse

Les techniques d'Internet ont largement évolué, ouvrant la voie à de nouveaux usages et de nouveaux comportements. Avec le Web 2.0 le mouvement du Web s'oriente vers plus de simplicité (aucune maitrise technique ni informatique n'est requise pour les utilisateurs) et d'interactivité (chaque internaute est invité, de façon individuelle ou collective, à participer , à partager et à collaborer sous différentes formes).

Le Web 2.0 suit la forme originelle du web 1.0, en particulier les interfaces qui permettent aux utilisateurs simples ayant peu de connaissances techniques de s'adapter aux nouvelles fonctionnalités du web. Ainsi, Le comportement des utilisateurs a totalement changé depuis l'apparition des applications Web 2.0 telles que les blogs, les wikis, les réseaux sociaux, les nombreuses plateformes de partage de contenus, etc. L'internaute devient le producteur des informations en interagissant de façon simple, à la fois avec le contenu et la structure des pages, mais aussi avec les autres internautes créant ainsi le Web social.

L'essor spectaculaire des sites collaboratifs a permis l'apparition de nouvelles formes d'indexations des contenus du Web créées librement par les internautes et

partagées au sein de réseaux sociaux, qui ont été baptisées du nom de folksonomies.

Les folksonomies résultent de la collection de tags partagés par les utilisateurs de plateformes de social tagging. Ces plateformes permettent à leurs utilisateurs d'organiser leurs resources favorites en leur associant de manière libre des signes appelées tags

Les folksonomies sont apparues depuis 2004 comme une alternative interessantes à des différentes politiques de structuration et d'accessibilité sur le Web, telles que les annuaires Web, les moteurs de recherche et les standards techniques proposés par le Web sémantique.

1.2 Problématique

Les tags sont des mots-clés, mais ils présentent un certain nombre de particularités. Les tags ne sont pas produits par des experts mais sont librement choisis par les internautes qui annotent des ressources mises en ligne.

Les tags ne sont pas considérés seulement comme des métadonnées isolées servant à faciliter la recherche des ressources, mais ils ont la particularité qu'ils forment un réseau connecté de tags produisant des liens entre les contenus et les utilisateurs qui les ont indexés. Ce qui représente une sémantique implicite avec un grand potentiel exploitable dans différentes application du Web telles que la recherche, les systèmes de recommandation et même l'amélioration du processus du tagging.

Si les folksonomies facilitent le partage et la collaboration entre individus, dans l'objectif de faire émerger une Intelligence Collective au sein du Web, elle introduisent de nouvelles problématiques en termes d'exploitation pertinente des informations produites. Ces problèmes dus notamment à l'ambiguïté et l'hétérogénéité des mots-clés utilisés, ainsi qu'a leur manque d'organisation.

Cette thèse s'inscrit dans le cadre des récentes recherches liées à la de convergence entre Web Sémantique et Web 2.0, deux approches du Web qui ont habituel-

lement été considérées, à tort, comme disjointes. Nous nous focalisons Notamment à la manière dont celles-ci peuvent cohabiter et bénéficier chacune des apports de l'autre. A cet égard, nous nous intéressons en particulier au rapprochement entre folksonomies et représentations structurées de connaissances tels que les theasauri ou les ontologies informatiques. Cela afin de tirer profit de la richesse des folksonomies et des interactions sociales issues de ces structures pour la représentation et l'exploitation de connaissances formalisées selon les principes du Web Sémantique.

1.3 Démarche de la recherche et contribution

Cette thèse propose donc une approche pour extraire la sémantique émergente dans les folksonomies. Ainsi, les contributions de cette thèse seraient les suivantes :

1. Une nouvelle representation de la folksonomy (agrégation) basée sur une matrice binaire tag-tag au lieu de resource- resource dans le but d'améliorer la performance des experimentations
2. Une nouvelle mesure de similarité plus simple et plus précise pour extraire les relations entre les tags
3. Une nouvelle mesure de généralité simple et efficace pour calculer le degré d'abstraction des tags, cette mesure est utilisée pour extraire la hierarchie des tags dans la folksonomie.
4. La prise en compte et la résolution du problème d'ambiguité lors du processus de generation de l'ontologie légère à partir de la folksonomie. Cela permet de retourner des resultats plus précis par rapport aux mecanismes de recherche traditionnels.
5. Nous tirons profis de la propriété du clustering flou qui permet au cluster de se chevaucher pour distinguer les tags ambigus et pour definir leurs différents sens à la fois.
6. Nous proposons un nouveau algorithme de clustering flou plus simple, plus efficace, et ne nécessitant aucune donnée préalablement definie comme le nombre des clusters à générer ou autre.

1.4 Organisation de la thèse

Cette recherche s'articule en quatre chapitres auxquels viennent s'ajouter cette introduction et une conclusion. En introduction nous avons déjà présenté la problématique et les contributions.

Le deuxième chapitre traite du contexte de la thèse caractérisé par l'apparition du web 2.0 et l'apport des utilisateurs sur ce web, l'apparition des systèmes de tagging collaboratifs ou les folksonomies et leurs caractéristiques, et enfin la complémentarité entre les folksonomies et les ontologies et à quel point cela pourrait servir pour combler leurs limites et tirer profit de leurs points forts.

Dans le troisième chapitre nous allons présenter les travaux connexes qui visent à extraire des relations sématiques à partir des folksonomies. Nous commoncerons par décrire de façon globale chaque approche en suivant une classification basée sur les techniques utilisées pour extraire ces relations. Ensuite nous projecterons ces approche sur un processus général afin de pouvoir les comparer.

Nous présenterons dans le quatrième chapitre notre approche en suivant le même processus, afin de mettre en valeur les améliorations qu'apporte notre approche par rapport aux autres citées précédemment. Une évaluation expérimetale de notre approche est faite dans le chapitre cinq afin de prouver l'efficacité et la performances des différentes contributions que nous avons proposées.

Le dernier chapitre, portant sur la conclusion générale de la thèse, présente nos conclusions et perspectives.

Chapitre 2

Background

2.1 Introduction

Dans ce chapitre nous décrivons le contexte dans lequel nos travaux ont été réalisés. Ce qui nous amène à nous intéresser aux spécificités de l'espace documentaire numérique que représente le Web et son évolution récente avec l'apparition du Web 2.0 auquel le développement du tagging est lié. Tout en présentant la complexité, la diversité et la particularité des folksonomies comme classifications générées par le tagging. Ensuite, comme solutions à ces difficultés, nous posons des hypothèses portant de créer une complémentarité entre ces données non structurées et les représentations structurées de connaissances telles que les theasauri ou les ontologies informatiques.

2.2 Historique du Web

Vers la fin des années 80, le Conseil Européen de la Recherche Nucléaire (CERN) hébergeait un grand nombre de chercheurs, provenant de différents pays, utilisant différents ordinateurs qui fonctionnaient avec des formats de fichiers différents. Ces chercheurs voulaient travailler ensemble, Mais l'échange des fichiers était très complexe suite à la nécessité d'effectuer des conversions entre différents formats. L'idée d'un espace informationnel commun s'imposait naturellement. Tim Berners Lee était à l'époque chercheur au CERN et il avait compris, qu'un tel moyen d'échanges d'informations, clairement nécessaire pour mieux connecter ses collègues au CERN, pourrait potentiellement connecter toute l'humanité.

A l'époque de la création du Web, l'Internet existait déjà. Mais si cette infrastructure permettait l'échange de fichiers elle n'était pas suffisante pour permettre à un chercheur de se servir d'informations produites par un autre chercheur à cause des formats de fichiers hétérogènes. Au lieu de faire une multitude de transformateurs de formats de fichiers, Tim Berners Lee envisagea un outil beaucoup plus puissant qui pourrait intégrer des informations produites par différents utilisateurs dans un espace informationnel commun, dans lequel celles ci pourraient être consultées, réutilisées et référencées.

Le concept d'hypertexte était également déjà connu à l'époque de la création du Web. Mais ces applications ont majoritairement été conçues pour fonctionner

sur un seul ordinateur ou dans un réseau local et n'avait pas la capacité de se connecter à l'Internet.

Tim Berners Lee voyait une opportunité dans la combinaison de l'hypertexte et de l'Internet. Il croyait que l'hypertexte était un moyen idéal pour interconnecter des informations de type différent, en approvisionnant une interface unique par différentes formes d'informations, telles que rapports, notes, bases de données, etc. [Berners-Lee, 1990]. C'est en envisageant un système hypertexte où les utilisateurs pourraient créer des liens entre des contenus résidant dans différents ordinateurs, et naviguer dans cet espace d'informations connectées à l'aide de l'Internet, qu'il a créé le Web [Berners-Lee, 2000].

Il est très important de souligner la distinction entre le paradigme du Web en tant que moyen de connexion entre un humain et le monde extérieur comprenant des mémoires, des expériences et d'autres humains, et la réalisation technique de ce paradigme. Si le paradigme est resté inchangé, l'implémentation technique du Web a évolué au fil du temps, notamment dans le sens de l'amélioration des standards, des protocoles et des technologies qui font vivre le Web. Cette évolution a eu une influence aussi bien sur la pratique de consommation du contenu Web, que sur la pratique de mise à disposition du contenu via le Web. Cette influence a également rendu l'usage du Web en général plus facile. Plus le Web devenait facile à utiliser, plus le nombre de ses utilisateurs grandissait. Plus le nombre de ces utilisateurs devenait important, plus il était économiquement intéressant de le faire avancer et le rendre encore plus utile. C'est ce cercle vertueux qui a conduit le progrès rapide du Web et sa transformation en une plateforme qui augmente les capacités cognitives de millions de personnes, dans une variété de situations quotidiennes.

Certaines améliorations de l'infrastructure du Web ont un impact tellement important sur la perception du Web par ses utilisateurs qu'elles créent l'impression d'avoir produit un nouveau Web. Cet effet a pu être remarqué pour la première fois avec l'introduction des sites Web qui permettent la création facile du contenu Web et la collaboration avec d'autres utilisateurs. Ce fut la naissance de Web 2.0, connu aussi sous le nom Web Social. Ensuite, l'initiative de rendre le contenu du Web plus structuré et plus manipulable par des machines fut appelée Web Sé-

mantique ou Web 3.0. Par Web des objets nous entendons souvent un Web où les appareils et d'autres objets physiques sont connectés et référençables sur le Web. Même si la multiplicité d'appellations des différentes phases et différents aspects du développement technique du Web ne semble pas converger vers une nomenclature commune, une catégorisation proposée par Nova Spivack [Spivack, 2009] s'est imposée par sa clarté et reste parmi les plus citées sur le Web. Cette catégorisation, présentée sur Figure 2.1 , explique les phases du développement du Web en fonction de leur apport en terme de la connectivité des personnes (sur l'axe X) et de la connectivité d'informations (sur l'axe Y) qu'elles apportent à la réalisation technique du Web. Cette perspective propose une vision de l'évolution du Web vue comme une oscillation permanente entre les directions de développement mettant un accent sur la connectivité des personnes, et celles mettant un accent sur la connectivité des informations.

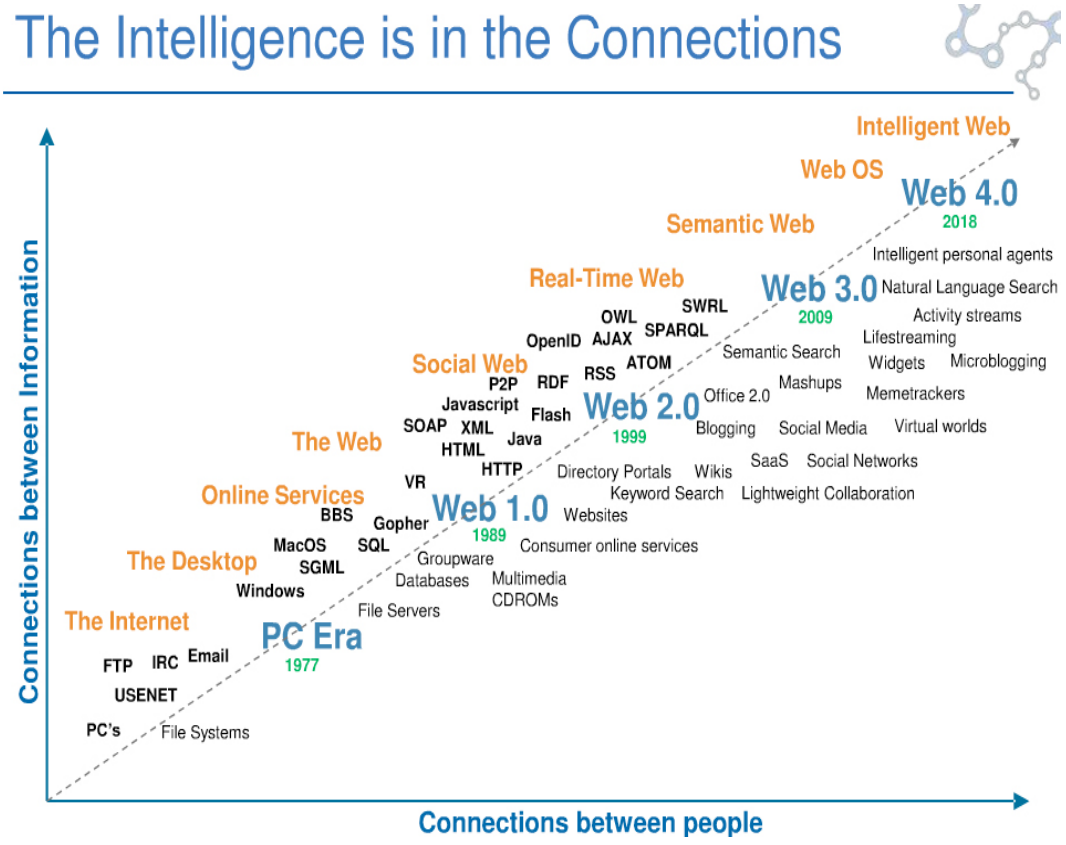

FIGURE 2.1 – Evolution du Web. Repris de Nova Spivack [Spivack, 2009]
2.1

2.3 Web 2.0

"*Le Web est une création sociale plutôt que technologique. Je l'ai créé en envisageant un effet social pour faciliter la collaboration entre des personnes, ce n'est pas un jouet technique.* " affirme Tim Berners Lee dans son livre Weaving the Web [Berners-Lee, 2000].

Malgré cette intention, dans les premières années du Web, il était difficile de se servir du Web pour communiquer. Le contenu du Web était créé principalement par une minorité d'utilisateurs formés au HTML, ce qui rendait le Web un canal de diffusion non participatif pour non initiés.

C'est avec l'émergence des blogs, wikis, forums et d'autre sites destinés à la collaboration que le Web a commencé à prendre la forme d'un outil d'intégration dans l'aspect social du Web. Ce phénomène et son impact sur l'industrie du Web et sur l'apparition de nouveaux services innovants ont notamment été repérés par O'Reilly en 2005 [Tim O'Reilly, 2005].

2.3.1 Les principes du web 2.0

Le texte de Tim O'Reilly, paru en 2005, sous le titre "Qu'est ce que le web 2.0 ?" [Tim O'Reilly, 2005] dégage sept principes clés du Web 2.0 (Figure 2.2). En voici un petit résumé :

1. **Le web en tant que plateforme :** Le Web devient une plate-forme pour la création de nouvelles applications c'est-à-dire qu'un site web peut être un outil aussi efficace pour certaines applications qu'un logiciel, tout comme le tableur et le traitement de texte en ligne offerts par Google.
2. **Architecture participative :**Les contributions des utilisateurs, sous forme de connaissances (ex. : Wikipedia) ou de commentaires et évaluations (ex. : Amazon.com), procurent une valeur ajoutée qui permet d'exploiter de manière efficace la force de l'intelligence collective potentiellement existant sur le web.
3. **La puissance est dans les données :**L'accessibilité des données permet la création de nouvelles applications combinant plusieurs sources de

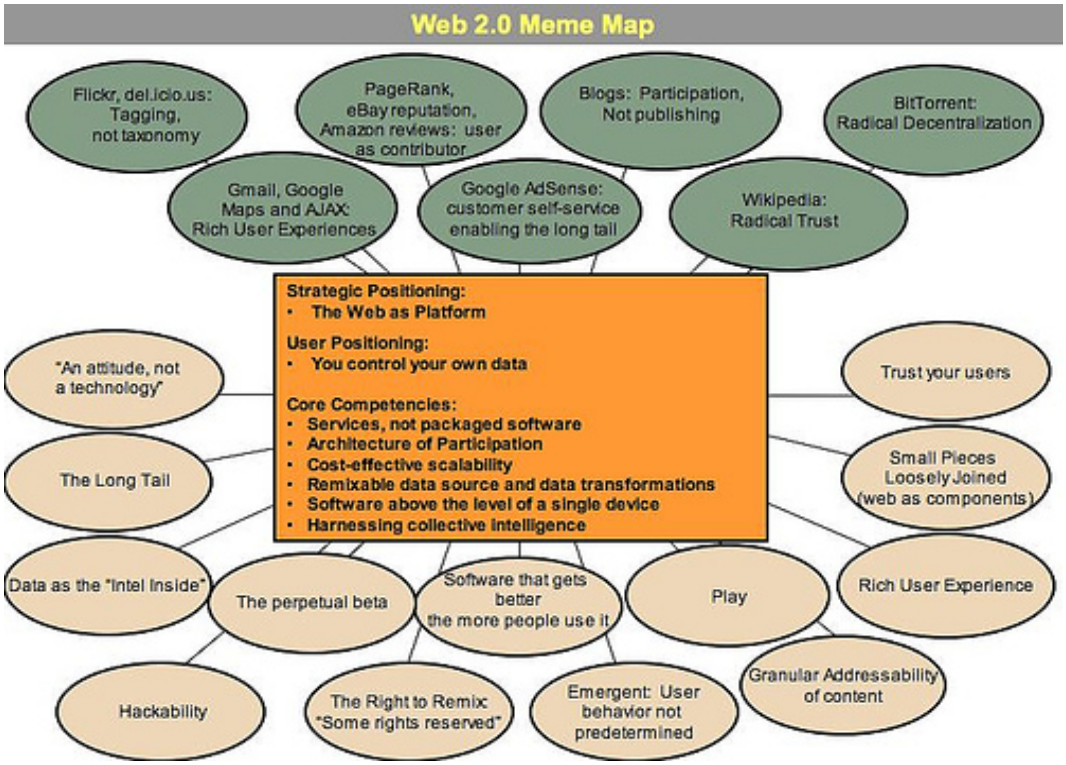

FIGURE 2.2 – Les principes du web 2.0
2.2

données qui sont les "mashups".

4. **La fin des cycles de release :** Les applications sont constamment améliorées et les changements sont disponible en ligne presque immédiatement, sans avoir à attendre les changements de versions (1.0, 1.1, etc.) comme dans les logiciels traditionnels. C'est le principe de la "version bêta perpétuelle".

5. **Des modèles de programmation légers :** Des technologies telles RSS et les services Web permettent la création d'applications faiblement couplées et plus flexibles.

6. **Le logiciel se libère du PC :** Une autre des caractéristiques du web 2.0 qui mérite d'être mentionnée est le fait qu'il n'est plus limité à la plateforme PC. Bien entendu, n'importe quelle application web peut être vue comme un logiciel indépendant d'une plateforme spécifique.

7. **Enrichir les interfaces utilisateur :** Des technologies comme AJAX permettent d'enrichir les interfaces disponibles sur le Web pour accomplir

diverses tâches.

Parmi cet ensemble de principes fondamentaux deux notions qui nous paraissent particulièrement importantes, à savoir celles du Web en tant que plate-forme et celle d'architecture participative (architecture of participation).

Cette première notion reconsidère l'utilisation duWeb et de ses principes pour y fournir des services et applications à forte valeur ajoutée plutôt que des contenus essentiellement statiques. Le rôle du Web peut même être dédié à celui de simple plate-forme d'échange et de transit de l'information comme dans le cas de RSS. Par extension, on regroupe également sous ce terme la migration de services traditionnels (client mail, suite bureautique ...) vers des applications en ligne.

Dans ce contexte, la notion d'architecture participative met en avant la production de contenus à forte valeur ajoutée par effet de bord des usages réguliers et des intérêts personnels que chacun poursuit en utilisant ces applications. Ceci se fait par ailleurs de manière autonome en raison de la manière même dont ces applications ont été conçues. Par exemple dans le cas des Wiki, des modifications établies individuellement permettent d'enrichir globalement un document ou un site de manière collaborative mais surtout continue et transparente.

O'reilly fait ainsi l'analogie avec les processus de développement open-source (le développement de fonctionnalités par un utilisateur pour un besoin précis impliquant une évolution générale de l'application dont tous peuvent bénéficier) et les architectures peer-to-peer (chaque consommateur devenant à son tour fournisseur de données) parviennent à ce même objectif. On peut également comparer ces principes à l'architecture même du Web, l'ajout d'hyperliens entre documents permettant d'accroître la structure du graphe global qu'il représente, renforçant ainsi les possibilités générales de navigation.

Plus généralement, on peut considérer le Web 2.0 comme une vision du Web mettant à disposition des utilisateurs un ensemble de services et de technologies visant à faciliter la production et le partage d'informations de manière intuitive et collaborative. Ainsi, le Web devient un média participatif many-to-many, plus

qu'un simple espace de stockage à titre essentiellement consultatif, l'utilisateur final ayant de ce fait un rôle central dans cette démarche. Pour y parvenir, les servicesWeb 2.0 partagent pour la plupart un ensemble de principes communs :

— l'utilisateur est au centre du service, en termes de publication et de réaction. On peut même aller jusqu'à dire qu'il fait l'outil, la valeur de ce dernier dépendant de son contenu. Nous nous situons ici dans un schéma inverse de celui des portails Web de la fin des années 90 abondés par une autorité ou une équipe de rédaction établie à priori. Nous pouvons ainsi considérer que de nombreux servicesWeb 2.0 sont des contenants vierges de tout contenu, ceux-ci étant soumis à l'adoption de l'outil par les utilisateurs ;
— le passage du statut de consommateur à celui de producteur doit se faire simplement. Le lecteur doit être en mesure de réagir à l'information qu'il consulte, a minima à un niveau inférieur à celui du producteur originel de l'information consultée (commentaires sur les blogs), au mieux au même niveau que celui-ci (édition de contenu sur un wiki, services de partage de contenu, etc.). Pour accentuer cette simplicité, les interfaces se doivent également d'être intuitives et sans prérequis technique ;
— la composante sociale se doit d'être présente non seulement en termes de publication mais aussi en termes d'échanges entre membres de la plateforme. De tels services doivent être en mesure de stimuler les synergies entre internautes, voire de participer à l'élaboration de réseaux sociaux, virtuels ou réels, certains outils y étant entièrement dédiés.

Malgré ces caractéristiques communes, les services offerts sont relativement divers. Ainsi, les blogs mettent en avant l'individu, en offrant un système de publication personnelle en ligne. Les wikis ont quant à eux pour objectif de participer à l'élaboration collective et consensuelle de contenu. De nombreux services de partage de données ont également fait leur apparition, favorisant généralement la définition de réseaux sociaux. Un autre aspect important dans l'utilisation de ces outils est la notion de mash-up, ou application composite, permettant de combiner les données provenant de divers services ou de les visualiser avec de nouvelles interfaces. De nombreux services Web 2.0 proposent en effet aux développeurs des APIs permettant de réutiliser les données produites en leur sein. Il

est ainsi possible de combiner les données provenant des différentes applications, mais également de les combiner avec d'autres interfaces de visualisation

2.3.2 Les services Web 2.0

Dans cette section nous verons en details les services cités précedement en explicitant la différence entre elles, et ses façon de collaboration qu'elles permettent aux utilisateurs.

2.3.2.1 Les blogs :

Un blog, diminutif de weblog, est un site présentant sur sa page d'accueil un ensemble de billets (posts dans le vocabulaire anglophone) consistant en des notes ou articles plus ou moins longs et ordonnés de manière antéchronologique, l'usage d'hyperliens (internes et externes) y étant abondant. Un blog est en général personnel et donc maintenu par un unique auteur ou blogueur, Mais peut aussi être partagé entre plusieurs rédacteurs, chacun ayant alors pour habitude de signer distinctement ses billets. En effet, le blog, contrairement au wiki que nous évoquerons dans la section suivante, met fortement l'accent sur la notion d'identité de l'auteur en tant que producteur de contenu. La notion de collaboration n'est alors pas liée à la rédaction de billets, mais à la possibilité que les lecteurs ont de réagir aux propos consultés par l'intermédiaire de commentaires associés aux billets. Cet aspect participatif permet ainsi à chacun de former et de fidéliser une communauté de lecteurs évolutive et réactive autour de soi et de ses écrits ou opinions.

À nouveau, ce n'est pas l'aspect technologique des blogs qui fait leur force, mais leur simplicité de mise en oeuvre et d'utilisation couplée à la composante collaborative évoquée ci-dessus. De nombreux services proposent la création d'un blog en quelques minutes (Blogger [1], Wordpress.com [2] ...) et les outils pour installer son propre système sont également nombreux. La publication se fait sans connaissance technique via une interface Web ou dans certains cas directement

1. https ://blogger.com
2. https ://wordpress.com

depuis son poste de travail ou un terminal mobile, contribuant à l'ubiquité de la présence en ligne d'un individu. Ainsi, les blogs ont remis au goût du jour le concept de page personnelle, la nature spontanée et régulière des billets et leur présentation antéchronologique offrant cependant une dynamique tout autre.

Une des forces des blogs, comme nous l'avons évoquée, est la possibilité d'expression spontanée qu'ils offrent et en conséquence les discussions qu'ils engendrent.

À cet égard, il nous semble important de signaler l'explosion récente du phénomène de microblogging, popularisé par Twitter [3]. À mi-chemin entre le blog et la messagerie instantanée, ce mode de communication se traduit par la publication de courts messages (généralement moins de 140 caractères) non-titrés et sans restriction de contenu. Si ces messages sont généralement proches de la notification de statut personnel, ils peuvent aussi servir au signalement léger d'informations (en postant par exemple un simple lien vers une ressource en ligne jugée intéressante) et permettent de manière plus générale une communication agile entre les personnes les postant et ceux y répondant ou simplement les suivant.

2.3.2.2 Les Wiki

Un wiki [Bo Leuf, 2001] est un site Web dynamique et évolutif, au sens où il permet à chaque lecteur de modifier les pages consultées et d'en ajouter de nouvelles mais aussi d'en supprimer. Ainsi, la dynamique d'un wiki s'observe non seulement vis-à-vis du contenu de ses pages mais aussi via l'architecture générale de celui-ci, évoluant selon les actions utilisateurs. Un wiki n'est généralement pas axé sur des informations contextualisées temporellement et produites par un auteur unique identifié (cas du blog), mais sur la construction collaborative et incrémentale de contenu consensuel. Les usages des wikis sont divers, de l'encyclopédie généraliste l'exemple le plus parlant étant Wikipedia [4], l'ouverture du site s'inscrivant ici dans la continuité du libre accès du code. Même si ces outils ont été popularisés récemment, le premier prototype de wiki date de 1994 [5] le nom trouvant son origine dans le terme hawaïen wiki wiki, signifiant vite. Parmi

3. http ://twitter.com
4. http ://www.wikipedia.org/
5. http ://trac.edgewall.org/

les caractéristiques essentielles des wikis, nous retiendrons :

— Des processus simples pour la participation. Par défaut, chaque lecteur doit être en mesure d'éditer le contenu d'un wiki quelque soit le niveau de modification souhaité (ajout, création ou suppression de pages) via le même outil que celui qui permet la visualisation du site,
— En conséquence de cette édition ouverte, chaque page doit bénéficier d'un historique des modifications. Celui-ci permet de revenir simplement à une version précédente (en cas de modifications jugées non souhaitées pour la communauté, ou de vandalisme) ou simplement de consulter les modifications apportées entre deux versions. Certains wikis permettent également de s'abonner au flux des modifications d'une page,
— Le rôle important joué par les hyperliens. Un wiki doit permettre d'établir facilement des liens entre pages du même wiki. Pour ce faire, on utilise généralement la syntaxe MotWiki qui permet d'établir automatiquement un lien vers une page portant ce nom ou d'en créer une si celle-ci n'existe pas. Cette pratique renforce la dynamique des wikis et évite la présence de pages orphelines, c'est-à-dire sans lien entrant. La notion de rétrolien est également très présente, chaque page listant l'ensemble des pages ayant un lien entrant vers celle-ci. Cette pratique étend ainsi la notion de source et de direction des hyperliens pour offrir une navigation à double sens entre les pages.

Si le principe d'ouverture des wikis en fait, dans l'idéal, un outil adéquat pour la constitution collaborative de documents ou de sites, il soulève de nombreuses questions et introduit également des problèmes de spam ou de vandalisme. Ainsi, si certains systèmes introduisent des restrictions d'accès pour la modification des pages, d'autres s'organisent comme des espaces autogérés où les utilisateurs rectifient eux-mêmes les pages modifiées dans un sens n'allant pas avec celui défini, explicitement ou non, par la communauté.

2.3.2.3 RSS

Devant cette abondance de contenus en ligne et leur régulière évolution, il est nécessaire de fournir un moyen d'obtenir le signalement d'informations pertinentes selon les centres d'intérêt de chacun. La syndication de contenu a pour

objectif de répondre à ce problème, en offrant aux sites un moyen de délivrer automatiquement un flux constamment actualisé de leurs dernières mises à jour, auquel les lecteurs peuvent s'abonner.Dans le but de formaliser ce processus et d'offrir un format standard de données, plusieurs modèles ont vu le jour, comme NewsML [6], dès 2000, pour les échanges entre fournisseurs d'informations et agrégateurs de données.
L'utilisateur peut souscrire à ces flux via un agrégateur, logiciel client ou service en ligne offrant une vision humainement lisible de ces informations brutes et tirant partie des différentes métadonnées contenues dans ces flux pour ordonner les éléments par date, source ou encore par auteur. Ces applications permettent également de récupérer à intervalles réguliers les dernières mises à jour des dits flux

2.3.2.4 Tagging et folksonomies

Enfin, face à cette abondance d'informations, facilitée par les outils et services présentés en amont, se pose le problème d'un accès pertinent à celle-ci. Jusqu'à présent, cette tâche était essentiellement rendue possible via des systèmes classiques d'indexation de pages Web. Le Web 2.0 a introduit une autre pratique, basée sur la catégorisation des contenus par les utilisateurs eux-mêmes via l'association aux ressources en ligne de mots-clés libres (aussi bien en type, nombre ou langue), ou tags. Il est important de noter que :

— d'une part cette pratique ne se limite pas aux données textuelles mais qu'il est possible de taguer des ressources numériques aussi diverses que des photos (Flickr) ou des vidéos (YouTube).
— d'autre part, certains sites proposent d'étiqueter non seulement les contenus des utilisateurs, mais aussi ceux, déjà tagués, d'autres utilisateurs (Delicious).

Cette pratique s'est également répandue sur la blogosphère, de nombreux billets de blog étant annotés de cette manière, un service comme Technorati permettant ensuite de visualiser ceux-ci et de restreindre la recherche d'information à un tag précis.

6. http ://www.newsml.org

De par son rattachement à un contenu existant, un tag peut essentiellement être vu comme une métadonnée supplémentaire associée à une ressource. Cependant, alors qu'un outil de blog associe automatiquement à un billet la date de création de celui-ci et le nom de son auteur, qu'une photo possède ses métadonnées EXIF pour identifier ses caractéristiques, les métadonnées ici générées sont de l'ordre de métadonnées contrôlées et personnalisées par l'utilisateur lui-même, ou métadonnées sociales. Si l'on se réfère à l'usage de métadonnées dans les bibliothèques numériques, on peut en identifier trois types [Taylor, 1999] :

— les métadonnées descriptives, caractérisant le contenu de la ressource et utilisées essentiellement dans une optique de recherche d'information
— les métadonnées structurelles, établissant des liens entre ressources et établies généralement de manière automatique depuis ces mêmes ressources ;
— les métadonnées administratives, qui définissent par exemple les droits d'accès ou les restrictions de copyright de la ressource.

Il est intéressant de constater que si la majorité des tags peuvent facilement être perçus comme des métadonnées descriptives (car essentiellement relatifs au contenu de la ressource, y décrivant ses sujets principaux), certains sont utilisés par les utilisateurs comme des métadonnées administratives ou même structurelles. Ainsi, on observe sur Delicious l'utilisation des tags *creativecommons* ou *gpl* relatifs aux licences du contenu annoté, ou encore *w3c* ou *slashdot* pour indiquer que la ressource est issue du site en question. Des études ont également montré que les tags pouvaient se révéler de diverses natures. Ainsi, Golder et al[Golder, 2006] ont identifié sept usages différents des tags comme l'annotation relative au contenu du document annoté (cas le plus classique), la référence personnelle (à lire), ou l'opinion au sujet d'une ressource (drôle). Marlow et al[Marlow, 2006] ont également montré que les tags pouvaient dans certains cas avoir un aspect social permettant à l'utilisateur de se mettre en avant (ex : vu en concert). Enfin, Berendt et al [Berendt, 2007] ont montré que les tags pouvaient dans certains cas, plus que des métadonnées, être considérés comme du contenu additionnel relatif à la ressource annotée. Quoi qu'il en soit, la pratique des tags est donc assez diverse est dépend fortement du contexte d'utilisation.

Il est important de souligner deux tendances qui émergent incontestablement sur le Web Social et qui transforment de manière significative la nature des éléments du Web :

— **La publication des éléments Web de petite taille**, appelés microposts. Les sites Web proposent de plus en plus des interactions nécessitant peu d'effort de la part de l'utilisateur. Par exemple sur le réseau social Facebook [7], l'utilisateur peut apprécier un objet Web en cliquant sur un bouton « j'aime » qui y est associé. Ces actions résultent de la création des nombreux micro-objets et intensifient l'interaction entre le Web et l'utilisateur. Un autre réseau social, Foursquare [8] permet aux utilisateurs de se déclarer présents dans un lieu localisé par GPS [NRC, 1995] en effectuant un « check in » qui sert à signaler leur présence dans ce lieu en temps réel à leurs contacts. L'émergence de tweets courts, faciles à créer, publier et republier fait également preuve de la réalité de cette tendance.

— **L'émergence du contenu en tant que conséquence indirecte d'une action de l'utilisateur.** Certaines applications Web, de plus en plus nombreuses, identifient les activités de l'utilisateur et génèrent du contenu Web, sans l'intervention de l'utilisateur. C'est le cas de Spotify, [9]le service de streaming musical, qui publie sur Facebook, de manière automatique l'information sur chaque chanson que l'utilisateur écoute.

Dans cette optique de croissance de la quantité et de la granularité d'informations publiées sur le Web, il est particulièrement important de pouvoir traiter et utiliser de manière pertinente toutes ces informations. La conscience de ce défis ainsi que de ces deux tendances ont imposé certains choix dans notre démarche scientifique. Elle peut, en partie, expliquer la nécessité d'une convergence du Web Social avec le Web Sémantique plus précisement entre les folksonomies et les ontologies. Dans ce qui suit nous essayons de voir en detail les folksonomies et ses caractéristiques.

7. http ://facebook.com
8. http ://foursquare.com
9. http ://spotify.com

2.4 Emergence des folksonomies et du tagging :

Les folksonomies représentent un des courants de politique de structuration des données à l'échelle du Web [Boullier, 2008], qui apparait comme une source de métadonnées susceptible de réconcilier les différentes politiques de structuration du Web. Dans cette section nous allons voir l'évolution de ces politiques qui permet l'apparition des systèmes de structuration à base de folksonomies.

2.4.1 Le Web des annuaires :

Les premières politiques de structuration du Web comportaient, pour chaque site, de créer une hiérarchie pour catégoriser les contenus. Il n'y avait pas de modèle général de la structuration des données du Web et chaque site créait sa propre classification des informations par les professionnels qui ont en charge sa maintenance. Ce modèle de catégorisation est encore existant sur le Web et les évolutions du Web jusqu'au Web 2.0 ne l'ont pas fait disparaître, mais dans un souci de clarté, il vise à limiter la présence d'une même ressource à l'intérieur de plusieurs catégories et les professionnels qui génèrent la classification hiérarchique sélectionnent les ressources de manière thématique. Les ressources indexées sont très limitées du fait du coût élevé de la tâche d'indexation.

2.4.2 Le Web des moteurs de recherche :

Face à la croissance exponentiel des sites Web et aux difficultés pour les utilisateurs de naviguer et de trouver les ressources qu'ils souhaitent à l'intérieur des annuaires, les moteurs de recherche sont apparus (Lycos, Alatvista, Google). Ces moteurs fonctionnent en explorant automatiquement le Web à l'aide de robots qui vont le parcourir pour en indexer le contenu. Le moteur de recherche permet de répondre aux requêtes des utilisateurs en fournissant un certain nombre de résultats, sous forme de liste de liens hypertextes. Les premiers moteurs de recherche comptent, à partir d'une requête d'un utilisateur, le nombre des occurrences des mots-clés recherchés, et l'ordre des résultats proposés aux internautes dépendait de la fréquence d'apparition de leur requête sur le site. Cette méthode est limitée car le nombre d'occurrences d'un mot-clé ne peut pas être considéré comme indice

de pertinence d'un site.

Une autre technique pour calculer la popularité d'un site a été tenue secrete par la société Google, avec son algorithme du « Pagerank » , cet algorithme a connu un succès important, car il garde la méthode du calcul du nombre des occurrences des mots clés recherchés et il l'améliore en prenant en compte la structure hypertexte des sites qu'il indexe. Ainsi, un site sera considéré d'autant plus pertinent qu'il est cité par d'autres sites qui proposent un lien vers lui. La popularité d'un site est donc calculée à partir de la structure du réseau de liens hypertextes qui permet d'observer les sites émettant un lien vers d'autres. Toutefois, tous les liens n'ont pas la même valeur pour l'algorithme, ainsi un lien provenant d'un site populaire n'a pas le même score qu'un autre provenant d'un site moins pertinent. De la même manière, l'algorithme HITS, se fonde également sur la structure du graphe hypertexte du Web pour calculer les scores des sites par rapport à une requête donnée et par la suite juger leur pertinence, en se basant sur la notion de « hub » sites qui ne contiennent pas nécessairement d'informations, Mais des liens vers d'autres, appelés « autorité », qui eux contiennent l'information , la pertinence d'un site est calculée par le nombre de leurs citations par les « hubs ».

Les moteurs de recherches ont donc fait passer le Web d'un modèle de structuration hiérarchique réalisé par des professionnels chargés de classer et de catégoriser les ressources web à une autre structuration automatisée reposant sur le contenu textuel des pages et sur le graphe hypertexte des sites indexés par les moteurs de recherche. Cela permet à l'utilisateur de parcourir un domaine beaucoup plus vaste, en lui présentant les moyens pour rechercher et évaluer des ressources relatives à sa requête.

Cependant, les moteurs de recherche soulèvent un certain nombre de problèmes [Crepel, 2011]. Tout d'abord, l'indexation des sites Web n'est pas exhaustive (couches bases du Web non accessibles, blacklistage qui supprime certains sites, etc.) et l'élaboration des algorithmes sur lesquels ils reposent ne sont pas neutres, mais se fondent sur des choix de développement qui vont avoir une incidence sur le type de résultats proposés aux utilisateurs, ils dépendent parfois également de logiques commerciales ou éditoriales, ce qui amène à valoriser cer-

tains sites dans les résultats de recherche.

2.4.3 Les folksonomies :

Avec l'apparition du Web 2.0, la structuration des ressources du Web a subi des changements importants. Les internautes sont impliqués dans la production des contenus, ainsi des quantités massives de ressources sont mises en ligne quotidiennement, donc l'indexation de ces ressources par les professionnels est devenue trop couteuse et difficile. D'autre part, la nature des ressources qui peuvent être des photos, des fichiers audio ou des vidéos complique la recherche de contenus pour les algorithmes qui reposent sur les données textuelles.

Les éditeurs de sites Web 2.0, notamment les sites de partage de contenus, vont donc présenter de nouveaux modèles de catégorisation qui se basent sur des classifications générées par les internautes afin d'indexer les contenus et les rendre accessibles. Ce modèle de structuration des ressources par les utilisateurs du Web s'appelle folksonomie (une catégorisation des ressources du Web par les internautes de manière individuelle ou collective).

2.5 Qu'est-ce qu'une folksonomie ?

Le terme de folksonomies désigne l'indexation communautaire, ou indexation sociale. C'est un néologisme inventé par Thomas Vander Wal, architecte de l'information, qui combine donc le terme taxinomie (règles de classification, taxonomy en anglais, qui est l'indexation traditionnelle professionnelle) et le peuple (folk). En d'autres termes, c'est la classification faite par les usagers.

Les folksonomies "constituent la possibilité pour l'usager d'indexer des documents afin qu'il puisse plus aisément les retrouver grâce à un système de mots-clés " [Deuff, 2007]. Elles sont l'application directe des principes du Web 2.0. En d'autres termes, les usagers, dans la bibliothèque 2.0, où se développent les folksonomies, ont un nouveau rôle, qui s'apparente à un nouveau pouvoir : celui d'indexer les ressources, comme le ferait un professionnel de l'information et de la documentation. Les folksonomies, par principe, n'ont pas de norme, elles sont ouvertes et sans contrainte. L'usager a une liberté totale lors du choix des tags qu'il va employer, cette indexation ne repose sur aucun thésaurus.

Les utilisateurs jouent, de cette façon, le rôle auparavant réservé aux documentalistes, dans le but de faire l'adaptation des langages documentaires et de l'indexation aux nouvelles technologies du Web.

Les internautes, dans l'usage de la folksonomie ne sont pas contraints à une terminologie prédéfinie mais peuvent adopter les termes qu'ils souhaitent pour décrire leurs ressources. Ces termes sont souvent appelés mots-clés ou tags.

2.6 Le tagging : Qu'est-ce qu'un tag ?

le « terme-clé », qui revient le plus souvent dans les folksonomies est celui de tag. Le tag peut désigner en fait un mot-clé, une catégorie ou une métadonnée. Le mot anglais tag signifie en français : étiquette de balisage, étiquetage, fléchage, marquage, voire traçage [Deuff, 2007].

Il s'agit d'un mot-clé ou terme associé à une ressource, qui sert à décrire telle ressource. Le tag se présente, donc, comme une métadonnée.

Les tags sont habituellement attribués de façon informelle et personnelle par les utilisateurs/internautes à leurs propres ressources. Le système de tagging ne fait pas partie d'un schéma de contrôle formellement défini.

Le tag peut alors prendre toutes les formes et les sens possibles, selon le désir de l'internaute et surtout selon sa culture et sa maîtrise de la langue. Comme le système de tagging ne s'appuie sur aucun thésaurus, des mots absents du dictionnaire ou des néologismes peuvent devenir des tags.

À l'utilisation de ces tags est liée la pratique d'étiquetage ou de tagging (figure 2.3) , association par un utilisateur d'un tag à une ressource donnée (billet de blog, photo ...). Cette relation qui forme ainsi une relation tripartite [Mika, 2007] peut se représenter par les trois éléments(Utilisateur,Ressource, Tag),telle que :

1. Utilisateur correspond à l'utilisateur qui effectue l'action
2. Resource correspond à la ressource annotée (billet de blog, page Web ...)
3. Tag correspond au tag utilisé

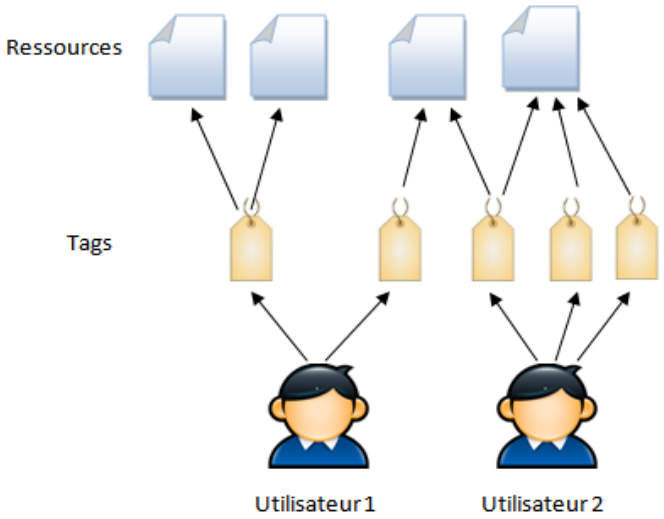

FIGURE 2.3 – Les éléments du tagging.

Les tags offrent la possibilité de rendre explicite les univers relatifs à l'information et de proposer de nouvelles formes d'accès et de mise en valeur des ressources Web à partir des métadonnées ainsi produites. Ils semblent pouvoir répondre en partie aux enjeux que représentent la structuration des données sur le Web. La particularité du mode de fonctionnement des folksonomies, qui se fonde sur des pratiques individualisées d'indexation de ressources documentaires numériques, nous amène à nous interroger sur la structure que prennent de telles formes de catégorisation et sur les aspects dynamiques de leur constitution.

2.7 Structure et distribution des folksonomies :

Les études qui portent sur les folksonomies s'intéressent pour la plupart à leur mode de structuration et aux formes de distribution des tags sur les sites Web qui intègrent des systèmes de tagging.
En se basant sur une analyse de données des sites Flickr (Figures 2.4 et 2.5) et Del.icio.us, M. Guy et al [Guy, 2006] s'intéressent à la distribution et à la popularité des tags sur ces sites Web. La distribution des tags suit une loi de puissance avec un phénomène de longue traîne, qui indique qu'une majorité d'internautes utilisent une quantité assez restreinte de tags différents, alors qu'une partie réduite d'entre eux utilisent un grand nombre de tags pour indexer leurs ressources. Ce phénomène de longue traine [Anderson, 2006] existe de manière

similaire dans les usages des plateformes du Web 2.0 et peut s'observer également sur la mise en ligne de contenu avec une majorité d'utilisateurs qui mettent en ligne une partie très réduite de contenus et une partie restreinte d'entre eux qui vont approvisionner le site d'une part importante de contenus. Les contributions des internautes à la constitution de la classification générale des contenus du site sont donc inégales.

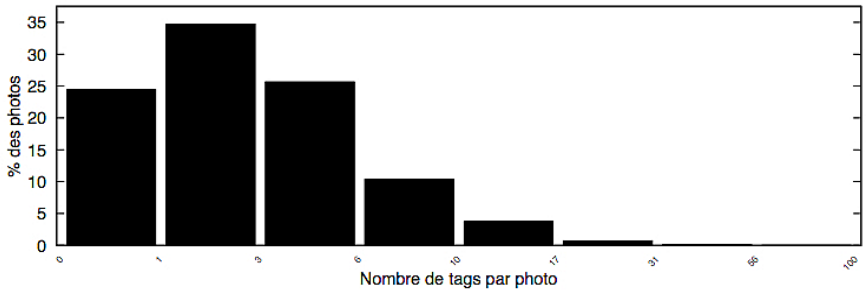

FIGURE 2.4 – Distribution des tags par photo sur Flickr [Guy, 2006].

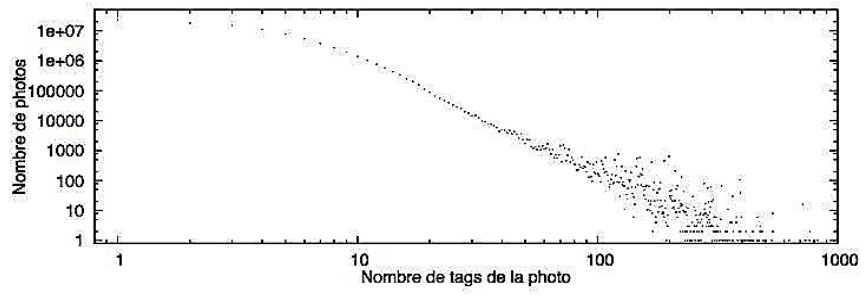

FIGURE 2.5 – Distribution du nombre de tags par photos sur la base de données [Guy, 2006] .

Les travaux de M. Guy et E. Tonkin [Guy, 2006] s'intéressent également à la nature même des tags postés par les utilisateurs et montrent que 10% à 15% des tags ne sont cités qu'une seule fois et que 28% sont mal orthographiés (du moins pas reconnus par des dictionnaires multilingues), ce qui amène à penser que le tags choisis par les utilisateurs relèvent de formes spécifiques, propres aux utilisateurs. De plus, les auteurs constatent une fréquente d'utilisation de symboles non textuels ou des tags sous des formes plurielles.

Une autre étude réalisée par [Spiteri, 2008] sur trois sites que sont Delicous, Furl et Technorati, compare les caractéristiques des folksonomies à un standard de construction de thesaurus nommé NISO [10]. Les folksonomies analysées dans cette étude comportent en moyenne 15% d'homonymes et seulement 4% d'erreurs orthographiques. Elles sont constituées en moyenne à 85% de mots uniques dont 80% sont des noms, 5% des formes verbales et 6% des adjectifs. Selon les sites on compte 3 à 10% de tags spécifiques, des formes de jargon ou de néologismes. Cette étude conclut que les caractéristiques des corpus de tags étudiés, à l'exception d'une présence plus importante d'homonymes, de synonymes, d'abréviations et de néologismes, restent en grande partie conformes aux standards de construction des thesaurus NISO et peuvent facilement être adaptés à des systèmes de vocabulaire contrôlé et normalisé produits par des professionnels.

Les travaux de Golder et al [Golder, 2006] s'intéressent quant à eux aux types de tags qui sont indexés par les utilisateurs et identifient sept types de tags différents pour définir les ressources sur le site Delicious. Les tags concernent :

— le contenu de la ressource (au sujet de quoi ou de qui)
— le type de support
— le contributeur ou la source du document
— une évaluation de la qualité du contenu
— une description des caractéristiques et un affinage de la définition du contenu
— la relation avec le tagueur (« mystuff »)
— une activité associée à la ressource (« jobsearch », « toread »)

2.8 Aspects dynamiqes des folksonomies

Il convient de considérer les folksonomies comme des classifications dynamiques, car elles évoluent au fur et à mesure de l'ajout de nouveaux contenus aux sites et des tags qui vont être indexés par les utilisateurs. Elle se construisent et se modifient à travers le temps mais sont également le fruit d'une agrégation des classifications personnelles d'utilisateurs, plus ou moins engagés dans l'activité, dont les schèmes classificatoires et les pratiques évoluent sans cesse selon

10. http ://www.niso.org/home/

leurs intérêts, leurs compétences et leurs appartenances.

Plusieurs travaux ont été effectués afin de comprendre la manière dont se construisent les folksonomies. Dans l'étude de Golder et al [Golder, 2006], qui se base sur des donnés issues du site Delicious, l'analyse de la répartition des tags dans le temps montre qu'il existe des phénomènes de saturation qui semblent indiquer l'émergence de conventions [Orlean, 1994] autour de la définition des ressources auxquelles ils sont indexés. Les fréquences de tagging les plus élevées s'observent au moment où la ressource est postée sur le site et connaissent généralement un niveau maximal au bout des six premiers mois. La particularité du système de tagging du site Delicious est qu'il permet à plusieurs utilisateurs de tagguer une même ressource non produite par l'utilisateur, en l'occurrence un site Web, ce qui permet d'observer la manière dont un même contenu va être indexé par plusieurs utilisateurs et d'analyser des phénomènes de convergences dans le choix des tags qui vont venir qualifier le contenu.

Cette émergence de conventions sur les modes de description à partir de tags communs et partagés semble aussi pouvoir être due au système technique qui propose aux utilisateurs des outils d'assistance au tagging basés sur la suggestion de tags qui ont déjà été indexés par d'autres utilisateurs, incitant de ce fait à mobiliser des tags déjà utilisés par d'autres. Au delà des outils de suggestion de tags pour assister les utilisateurs, il semble que l'interface du site Delicious, en rendant visibles les tags précédemment indexés par d'autres utilisateurs, tend à produire des phénomènes d'imitation, ou d'influence sur les choix des catégories mobilisées par l'utilisateur pour indexer la ressource [Cattuto, 2008].

Ce phénomène d'imitation largement mis en évidence sur Delicious permet de comprendre la dynamique de construction des folksonomies et le comportement des utilisateurs face ce système de catégorisation. Cette tendance à la convergence des tags par un processus d'imitation dans le choix des catégories d'indexation d'une ressource permettrait, à partir de modèle de simulation, d'anticiper le comportement des utilisateurs et d'inférer sur les tags cooccurrents les plus probables, permettant ainsi d'affiner les outils de suggestion et d'assister les utilisateurs dans leur tâche d'indexation [Zhang, 2009].

Les résultats de ces différents travaux nous apprennent que les folksonomies ne sont pas un système totalement désorganisé dans lequel les utilisateurs indexent de façon anarchique les ressources en ligne mais sont régulés en partie par les systèmes techniques et par les usages qui en sont fait.

2.9 Intérêt et usage des folksonomies

L'intérêt des folksonomies est lié à l'effet communautaire : pour une ressource donnée sa description est l'union de l'ensemble des descriptions de cette même ressource qui ont été faites par différents internautes. Ainsi, partant d'une ressource, et suivant de proche en proche les termes (tags) choisis par des autres contributeurs, il est possible d'explorer et de découvrir des ressources connexes et liées.

L'ensemble des mots-clés d'une personne peut être visualisé par des nuages de mots clés. Ce concept permet un survol de l'ensemble des centres d'intérêts d'une personne ou même d'un groupe. Les nuages de tags « tagcloud » sont une sorte de condensé sémantique d'un document dans lequel les concepts clés sont dotés d'une unité de taille (dans le sens du poids de la typographie utilisée) permettant de faire ressortir leur importance : les plus gros sont les plus utilisés par la personne/communauté.

Del.icio.us, Flickr, Box et Wikipedia sont respectivemen des exemples de site web qui utilisent les folksonomies comme sites de partage de signets/URLs favoris, de partage de photos, de stockage en ligne et comme encyclopédie collaborative.

Thomas Vander Wal distingue deux types de folksonomies, les « étroites » (narrow folksonomies) et les « générales » (broad folksonomies). Les folksonomies étroites sont surtout utilisées dans un objectif individuel tandis que les générales privilégient l'aspect collectif et collaboratif du partage d'information. Ainsi les sites de partages de favoris, comme del.icio.us ou Connotea, sont plutôt des folksonomies générales puisqu'un même site peut être partagé par plusieurs utilisateurs et recevoir le même tag. Ce type de folksonomie s'appuyant sur des réseaux sociaux ne fait pas que classer de l'information et la partager. Il met en

relation des utilisateurs qui partagent les mêmes centres d'intérêts. L'utilisateur-indexeur devient à son tour un peu indexé et mis en relation à la fois avec d'autres mots-clés, d'autres sites et d'autres usagers.

2.10 Avantages et Inconvenients des folksonomies :

Les folksonomies sont souvent comparées aux modèles de structuration traditionnels. Habituellement, l'indexation des contenus est réalisée par des professionnels (libraires, catalogueurs, bibliothécaires, développeurs, etc.), alors que dans le cas des folksonomies, cette indexation est réalisée par les utilisateurs. Ce modèle de structuration basé sur la participation des utilisateurs est remis en cause car, les folksonomies sont constituées d'un important bruit dans la classification, caractérisé par des phénomènes de synonymie, d'homonymie, d'anomalies lexicales ou encore par des pratiques « d'infopollution » qui tendent à piéger les utilisateurs qui recherchent des ressources documentaires. De plus, traditionnellement le coût important du travail d'indexation effectué par les professionnels est justifié être au bénéfice des utilisateurs qui ne peuvent avoir confiance qu'à une classification précédemment établie par des professionnels et qui facilitent leurs recherches sur le web. Cependant, ces problèmes des systèmes de tagging ne les ont pas empêchés de réaliser un grand succès et un essor spectaculaire sur le Web, cela revient aux points suivants :

1. Le tagging collaboratif réduit le coût d'indexation, en partageant la tâche d'indexation entre les utilisateurs.
2. Contrairement au modèle de structuration traditionnel, les folksonomies n'obligent pas les utilisateurs à s'adapter aux catégories délimitées a priori par les professionnels mais elles leur permettre de catégoriser selon leur propre système de catégorisation et leur logique d'association conceptuelle.
3. L'abondance des synonymes pourra servir à catégoriser une ressource en s'appuyant sur les différents sens qu'elle a. Les ambiguïtés sémantiques que produisent les folksonomies, du fait de la diversité des contributeurs qui les ont indexées, peuvent être considérées comme un avantage pour l'adaptation de la classification à un spectre large d'utilisateurs.

4. Elles présentent un certain nombre de propriétés avantageuses pour la navigation sur le Web. En effet, elles favorisent le principe de « sérendipité », autrement dit, la découverte inattendue de ressources, par la possibilité qu'elles offrent, de proposer des modes d'accès transversés à l'information et par les ambiguïtés sémantiques qu'elles produisent ([Quintarelli, 2007], [Weinberger, 2007]).
5. Contrairement aux catégorisations traditionnelles souvent trop réductrices , difficilement modifiables, et n'ayant de pertinence que dans des communautés restreintes d'experts, les folksonomies peuvent prendre en compte la diversité des langues, des points de vue et leurs évolutions dans le temps. ([Shirky, 2007], [Weinberger, 2007]).
6. Bien que les folksonomies soient produites, le plus souvent, dans une logique individuelle de classement de l'information où les tags postés par les utilisateurs sont produits de façon individuelle, elles renvoient systématiquement à des concepts partagés au sein de groupes sociaux qui leur donnent leur pertinence et leur légitimité.

2.11 Convergence entre ontologie et folksonomie

Les systèmes basés sur les folksonomie ont récemment tenté de répondre aux différents problèmes soulevés pour partager et indexer le grand nombre de ressources disponibles sur le Web. D'autre part, le Web sémantique vise à faciliter l'échange d'information en permettant l'interopérabilité entre les applications disponibles sur le Web. À cette fin, plusieurs méthodes, outils et principes sont proposés, parmi lesquels les ontologies formelles qui jouent un rôle central.

D'une manière générale, Ontologie est le terme utilisé pour désigner "une compréhension partagée d'un domaine donné" [Guarino, 1993]. Guarino établit un aperçu compréhensible de la définition d'ontologie à partir des travaux les plus cités de la communauté d'ingénierie des connaissances. La définition de Gruber [Gruber, 1993] est celle qui est la plus citée : "une ontologie est une spécification explicite et partielle rendant compte d'une conceptualisation". Pour Borst [Borst, 1997], en modifiant légèrement la définition de Gruber : "l'ontologie est

une spécification formelle d'une conceptualisation partagée.

Plusieurs éléments ou composants constituent une ontologie. Ceux qui reviennent le plus dans la littérature sont, (1) les concepts (souvent représentés par des termes), (2) les relations entre ces concepts (telles la relation sous-classe-de ou encore partie-de), (3) les fonctions, qui sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 premiers, (4) les axiomes, utilisés pour structurer des "phrases" qui sont toujours vraies et (5) les instances, utilisées pour représenter des éléments.

On distingue des ontologies de différents niveaux de généricité : des ontologies dites de haut niveau qui contiennent « des concepts très généraux comme l'espace, le temps, la matière, les objets, les événements, les actions, etc. [qui] ne dépendent pas d'un problème ou d'un domaine particulier » [Lando, 2006] ; des ontologies de domaine (médecine, architecture, mécanique..) ; de tâche (diagnostiquer, enseigner) ; voire d'application, dans lesquelles les concepts appartiennent à un domaine et à une tâche particulière (enseigner la médecine). Les ontologies ont plusieurs caractéristiques importantes :

1. Comme d'autres langages de représentation des connaissances, elles n'ont pas une vocation exclusivement documentaire au sens de l'indexation et de la recherche d'information mais elles visent aussi à participer de l'ingénierie des connaissances d'un domaine et en particulier à « spécifier explicitement une conceptualisation » pour reprendre les termes de T. Gruber.
2. Point corollaire du précédent, elles n'ont pas à être conçues strictement à partir d'un fonds documentaire qu'elles viseraient à indexer. Même si les ingénieurs de la connaissance responsables de leur conception utilisent souvent des outils terminologiques appliqués à des textes de référence du domaine, ils peuvent également s'appuyer sur d'autres sources d'information comme des entretiens auprès d'experts, l'analyse de bases de données, ou des conceptualisations ad hoc issues de leur propre synthèse des connaissances du domaine considéré.
3. Bien que la dimension documentaire ne soit pas, comme nous venons de le mentionner, leur justification intrinsèque, leur ancrage au sein du web sémantique tendrait néanmoins (pour autant qu'elles se diffusent effectivement) à leur faire jouer un rôle essentiel dans la recherche et la mise

en relation d'information. Mais l'information dont elles visent à faciliter l'accès est d'abord celle du web invisible, celui constitué par les multiples bases de données qui consignent l'information structurée des processus d'affaires et des références techniques. Cette vocation première est aujourd'hui concurrencée par l'usage des ontologies pour annoter des documents plus classiques. Mais cette tendance reste minoritaire.

4. En conséquence, les ontologies formelles ne sont pas faites pour être directement exploitées par des usagers humains engagés dans une navigation hypertextuelle comme cela pourrait être le cas pour une classification documentaire ou un thésaurus. Au contraire, elles sont le plus souvent conçues pour être exploitée par des programmes informatiques (des agents de recherche automatique sur le web), l'utilisateur interagissant avec l'agent à l'aide d'un formulaire ou d'un autre type de langage de requête.
5. De ce fait, les ontologies gagnent à être représentées à l'aide de langages formels, le standard proposé par le W3C (World Wide Web Consortium) étant aujourd'hui OWL (Ontology Web Language), qui s'exprime à partir du langage RDF (Ressource Description Framework), proche des réseaux sémantiques, lui-même exprimé à l'aide de balises XML comme tous les langages du web sémantique. Les classifications exprimées en OWL s'appuient sur une stricte séparation classe/instance, l'héritage de propriétés, l'expression de contraintes de cardinalité et de contraintes logiques sur les relations entres propriétés, etc. Cette formalisation extrême, vise à répondre aux objectifs ultimes du web sémantique tel que définis par T. B. Lee qui était de fournir des réponses logiquement fondées (« vraies ») aux requêtes des utilisateurs.
6. Enfin, la raison d'être première des ontologies formelles, liée à la manipulation des données structurées du web invisible en réponse à des requêtes complexes sur la base d'une sémantique formelle, a des conséquences sur le type de signification associé à ces langages. La sémantique des ontologies est une sémantique référentielle au sens du positivisme logique, les termes recevant une valeur de vérité ancrée sur des référents externes objectivables. Les concepts décrits par les termes de l'ontologie ont donc également principalement une valeur référentielle comme dans la tradition artistotélicienne où « le sens d'un signe est conçu comme représentation mentale (concept), et défini par ce à quoi il renvoie dans le monde (le

mot « chien » « signifie » ce quadrupède à poils ras). ». Or, cette vision du concept est largement incompatible avec les épistémologies de la philosophie pragmatique (J. Dewey) ou de la tradition herméneutique (cette dernière étant largement répandue dans les sciences humaines et sociales), épistémologies que nous défendons dans le cadre du web socio sémantique.

plus spécifiquement, les ontologies formelles utilisent la sémantique formelle pour spécifier cette conceptualisation et la rendre traitable par les machines. La généralisation d'utilisation des ontologies trouvent des difficultés principalement au niveau du coût de leur conception et leur entretien.

Les folksonomies s'appuient sur des dispositifs informatiques donnant « la possibilité à l'usager d'indexer des documents afin qu'il puisse plus aisément les retrouver grâce à un système de mots-clés » [Deuff, 2007]. Chaque utilisateur des plateformes telles que Del.icio.us [11] ou Flickr [12], peut déposer des ressources, marque-page ou photo personnelle, et leur associer des mots-clefs qu'ils peut ensuite partager avec les autres utilisateurs. Malgré les défauts liés à la faible cohérence des descripteurs (synonymie, polysémie, non explicitation des facettes prises en compte, absence de relation sémantique) les folksonomies semblent connaitre un réel succès. Comme le soulignent O. [Ertz, 2006]), les folksonomies, qu'ils assimilent à des pratiques d'indexation sociales, tirent leur force de deux phénomènes. D'une part, du faible effort cognitif requis par leur utilisation en comparaisons des classifications épistémiques de la bibliothéconomie et, d'autre part, de la fonction de régulation offerte par la mise en visibilité des mots-clefs déposés par l'ensemble des utilisateurs qui permet d'avoir un effet de feed-back rapide sur leur popularité et leur degré de couverture [Ertz, 2006]. Cet effet est renforcé par la possibilité d'accéder directement au site identifié par le marque-page (Del.icio.us) ou à la photo indexée (Flickr), ce qui réduit le cout potentiel de l'erreur d'indexation et permet de désambigüiser rapidement certains mots-clefs. Ces propriétés sont d'autant plus essentielles qu'à la différence des annuaires de ressources Internet qui sont contrôlés par des indexeurs sélectionnés selon leur réputation, les folksonomies sont totalement ouvertes au public inscrit sur les sites.

11. https ://delicious.com/
12. http ://www.flickr.com

Les folksonomies, ne représentent pas des innovations majeures du point de vue de l'organisation conceptuelle des descripteurs. Elles correspondent à des listes de vocabulaire dont la cohérence apparait comme bien faible eu égard de celles proposées par les professionnels de la documentation. L'innovation majeure se situe dans le processus collaboratif de construction des schémas de classification ou des listes de descripteurs et dans le processus d'indexation associé à cette construction à partir d'un flux de documents primaires très hétérogènes et dont le volume s'accroit très rapidement.

Dans la perspective du développement du web socio-sémantique, mais également dans la perspective du développement de nouveaux outils de gestion des bibliothèques numériques, l'hybridation entre les ontologies et les folksonomies, nous semble très prometteuse. Dans cette thèse, nous essayons de montrer le potentiel de combiner les deux approches afin de d'élaborer une représentation solide des connaissances qui sont à la fois représentatives des communautés d'utilisateurs, et qui permettent en même temps une meilleure extraction et échange d'informations

2.12 Conclusion

Le terme de folksonomie est apparu récemment sur le web pour désigner le phénomène d'indexation des documents numériques par l'usager [Deuff, 2007]. Le choix de ce mode de classificarion plutôt qu'un autre aboutit en réalité des contraintes spécifiques, et en retour, offre des possibilités nouvelles aux utilisateurs. Étudier le processus qui est à l'oeuvre dans ce type de catégorisation des données et les implications probables qu'il peut avoir pour les utilisateurs du Web, consiste à comprendre les avantages et les inconvénients de l'utilisation d'un mode de classification libre et subjectif plutôt qu'un mode de classification préétabli par des spécialistes. Comme nous l'avons remarqué, les folksonomies et les usages qui y sont liées, du fait de leur émergence récente ne sont pas normalisées et représentent un objet de recherche stimulant pour saisir la dynamique.

Dans ce chapitre, nous avons défini le tagging et les folksonomies, présenté les caractéristiques structurelles de ce mode de catégorisation des données et exposé les controverses qu'ils suscitent. Nous avons constaté que la meilleure solution

à cette difficulté de choix entre les deux modes d'indexation est de combiner les deux pour tirer profit de leurs avantages et remédier à leurs inconvénients. Dans le chapitre suivant nous présenterons quelques approches visant à rapprocher le mode d'indexation effectuée par les usagers des systèmes d'étiquetage collaboratifs, et le mode basé sur des experts.

Chapitre 3

Des folksonomies aux ontologies : etat de l'art

3.1 Introduction

Plusieurs pistes de recherche ont comme objectif de faire évoluer les systèmes basés sur des folksonomies en profitant des métadonnées produites par les utilisateurs afin de proposer des systèmes d'indexation, de navigation et de recherche innovants et de contribuer à l'émergence d'un Web plus sémantisé. ces recherches tendent à détacher les folksonomies de leur forme originelle qui est le tagging libre des ressources par les utilisateurs, soit en tendant de normaliser les usages et les tags eux mêmes, soit en utilisant les corpus de tags pour reproduire des formes de classification plus structurées. [Crepel, 2011] a différencié quatre orientations que nous allons présenter dans ce chapitre ensuite nous nous focalisons sur l'axe de recherche qui correspond à notre travail.

3.2 Les différentes pistes de recherche tentant de faire évoluer les folksonomies

Après s'être principalement attaché à comprendre les systèmes de tagging, leurs spécificités et les différences avec les formes antérieures de catégorisation de documents, la majeure partie des travaux de recherche s'est concentrée sur les questions de structuration des folksonomies. L'abondance de ces récents travaux de recherche démontre le potentiel que détiennent les folksonomies comme apport à la structuration des ressources du Web.

Les différentes pistes de recherche que Crepler [Crepel, 2011]présenté sont :

— Le traitement sur les corpus de tags a posteriori pour les rendre plus cohérents et réintroduire des relations entre les tags dans les folksonomies.
— Les projets de standardisation des tags pour les adapter aux formats du Web sémantique et à ses ontologies.
— Les approches hybrides articulant folksonomies avec d'autres systèmes de catégorisation plus structurés (thesaurus, vocabulaire contrôlé) comme dans le projet du Web socio sémantique.
— L'assistance par la suggestion et le tagging automatisé pour canaliser les usages et apporter plus de cohérence dans l'indexation des ressources à

partir des tags.

Notre travail s'inscrit dans la première piste qui sera présentée dans la section suivante. Dans cette section nous survolerons les autres axes de recherche et nous citerons quelques exemples des travaux inhérant à chaque axe.

3.2.1 Normalisation des usages du tagging : standardisation des tags pour le Web sémantique

Cette approche tend à enrichir les systèmes de tagging actuels en mettant à disposition des outils techniques d'annotations davantage normalisés et structurés afin d'orienter les utilisateurs à structurer les métadonnées qu'ils produisent. Différentes tentatives de standardisation des tags sous formes de triplets standardisés semblables à des tags RDF proposées dans le projet du Web sémantique existent déjà. On peut citer par exemple :

— Common Tag [13]
— SCOT [14] : Social Semantic Cloud Of Tags [Kim, 2008]
— MOAT [15] : Meaning Of A Tag [Passant,2008]
— NAO [16] : Nepomuk Annotation Ontology
— Newman's Tag ontology [17]
— TagOntology [18] [Knerr, 2007]
— UTO [19] : Upper Tag Ontology [Ding, 2009]
— Nicetag [20] [Gandon, 2009]
— Machine Tag (Flickr) [21]

Ces différents standards proposent un langage de description des tags qui permettent de lever les ambiguïtés et les incertitudes liées à l'utilisation d'un tag comme descripteur d'une ressource en enrichissant le modèle à partir du format

13. http ://commontag.org
14. http ://scot-project.org
15. http ://moat-project.org
16. http ://www.semanticdesktop.org/ontologies/nao
17. http ://www.holygoat.co.uk/projects/tags/
18. http ://tomgruber.org/writing/tagontology.htm
19. https ://scholarworks.iu.edu/dspace/handle/2022/9975
20. http ://ns.inria.fr/nicetag/2009/09/25/voc.html
21. http ://code.flickr.com/blog/2008/12/15/machine-tag-hierarchies/

RDF issu du Web sémantique. Ainsi à chaque tag est associé, sous forme d'un graphe nommé, une série d'informations qui concerne par exemple l'identité du taggueur, la ressource associée au tag, la date de tagging, les descripteurs qui permettent d'associer le tag à une classe plus large de concepts, etc. L'objectif de ces différents standards est de permettre une compatibilité entre le tagging libre et les standards de la W3C, fondés sur les ontologies. Dans une optique similaire, le développement de microformats, tels que « Xfolk » [22] par exemple, serait susceptible de permettre une indexation des pages Web par des tags lisibles, à la fois par les utilisateurs, mais aussi par les machines, en intégrant des tags au code HTLM des pages Web, sans pour autant nécessiter un reformatage complet des sites. Ce type de développement permettrait d'affiner l'indexation des contenus de pages Web et serait plus efficace, de ce fait, pour faciliter le travail des moteurs de recherche.

Cependant, la multiplicité de ces projets démontre la difficulté d'adopter un standard unique et interopérable. La démarche proposée, si elle permet une ouverture du modèle du Web sémantique vers des ontologies plus souples, s'adresse pourtant à un public très restreint de développeurs maîtrisant l'usage du code informatique et non pas à la masse des utilisateurs qui utilisent aujourd'hui le tagging sur le Web, et permettent l'enrichissement des bases de données par leur contribution à l'indexation des ressources. Enfin, par leur volonté de rendre compatible utilisation des tags et standards du Web, ces projets opèrent un choix fort, consistant à faire évoluer le tagging vers une plus grande structuration.

3.2.2 Systèmes de désambigüisation à partir de ressources externes et Web socio-sémantique

Le projet du Web socio-sémantique [Zacklad, 2007] propose de développer des outils de collaboration basés sur un système d'ontologies « sémiotiques », également appelées ontologies « légères », plus souples que les ontologies « formelles » sur lesquelles le projet du Web sémantique, dans sa version classique, se fonde. Nous pouvons considérer cette approche comme une étape intermédiaire entre le modèle prôné par le Web sémantique et celui du Web 2.0.

22. http ://microformats.org/wiki/xfolk

Le Web socio-sémantique, en intégrant différents systèmes de structuration et d'accès aux données, tente d'intégrer les différents politiques, en produisant des interfaces hybrides permettant à l'utilisateur de mobiliser différents systèmes selon ses besoins dans une démarche de recherche ouverte d'information.

Il existe également de nombreux travaux de recherche qui développent des systèmes de suggestion et de désambigüisation en associant folksonomies et systèmes de classification des données structurées par des autorités tels que des thesaurus ou des taxinomies disponibles sur le Web, en introduisant des modules implémentés aux systèmes de tagging qui viennent assister l'utilisateur dans sa tâche d'indexation [Laniado, 2007], [Passant, 2007].

De nombreux autres travaux utilisent des méthodes semblables de désambigüisation pour produire des interfaces hybrides intégrant tagging libre et bases de données structurées. Nous pouvons citer par exemple différents projets tels que :

— Facetag [Quintarelli, 2007]
— Semkey [Marchetti, 2007]
— T-org [Abbassi, 2007]
— Tagpedia [Ronzano, 2008]
— SRTag [Limpens, 2010]

3.2.3 L'assistance par la suggestion

D'autres outils d'assistance au tagging, déjà développés sur le Web, proposent des modules de suggestion de tags à partir des folksonomies constituées par les utilisateurs des sites, en se basant sur une analyse des cooccurrences de tags indexés. Ces outils d'assistance développés dans les systèmes de tagging proposent de fournir des suggestions, de grouper certains tags en les associant, de qualifier ou de décrire de manière plus approfondie le sens qui leur est donné.
Ces systèmes de suggestion de tags représentent également une tentative de normalisation des folksonomies, a priori, en s'appuyant sur des outils techniques d'assistance qui tendent à mieux structurer les folksonomies et les choix des tags par les utilisateurs. De tels suggestionneurs de tags existent déjà sur des sites tels que Delicious ou Flickr.

Après avoir discuté les différentes pistes de recherche ayant comme objectif de faire évoluer les folksonomies, nous allons nous focaliser sur le domaine de notre recherche qui vise à extraire les relations sémantiques entre les tags de la folksonomie.

3.3 Traitement des folksonomies pour extraire des relations entre les tags

De nombreux travaux poursuivent l'objectif d'extraire des modèles structurés, taxonomies ou ontologies, depuis les folksonomies, principalement dans l'objectif de résoudre les problèmes classiques des systèmes à base de tags tels que la synonymie et l'ambigüté. L'objectif est alors d'expliciter la sémantique qui peut exister dans ces systèmes là où celle-ci n'est qu'implicite en raison de la nature même des folksonomies. Dans ce chapitre, nous explicitons les différentes approches existantes visant à extraire la sémantique des folksonomies. Nous avons classé ces approches selon quatre axes : les approches basées sur les techniques de clustering, les approches basées sur les techniques de data mining, notamment les règles d'association, les approches basées sur l'alignement sémantique, et les approches hybrides. Dans ce qui suit, nous présentons ces travaux chacun dans sa catégorie.

3.3.1 Approches basées sur les techniques de clustering

Ces approches identifient la sémantique des tags en les regroupant dans des clusters selon certaines relations entre eux. Mika [Mika, 2007] considère un graphe triparti constitué des utilisateurs, des tags et des contenus. Ce graphe permet de générer des ontologies et de les faire évoluer de façon dynamique, en regroupant les tags à partir de leur inclusion dans des communautés d'utilisateurs, afin de former des clusters cohérents.

Hamasaki et al. [Hamasaki, 2007] ont étendu les travaux de Mika en introduisant la notion de voisinage des utilisateurs, en particulier, l'ontologie est modifiée en prenant en compte les informations des voisins des utilisateurs de la folksonomie.

Heymann et al. [Heymann, 2006] ont proposé un algorithme pour générer des structures hiérarchiques à partir des tags. L'algorithme se base sur le degré de similarité entre les tags, et le degré de généralité entre eux. Une extension de cet algorithme a été proposée par Benz et al. [Benz, 2010], où les auteurs introduisent des tâches d'identification de contexte et de désambiguïsation à l'algorithme.

Dans [Marouf, 2013], nous avons proposé une nouvelle mesure de similarité appelée CDU (co-occurrences par utilisateurs distincts), qui exploite les trois modes de la folksonomie. Après le nettoyage des données, les tags sont représentés dans un espace de vecteurs. Ensuite nous calculons la matrice de similarité entre les tag en employantla mesure Cosine. Un algorithme de clustering flou FCM [Bezdek, 1981]) est effectué sur la matrice Cosine pour assurer l'identification du Contexte des tags. Le Contexte d'un tag est l'ensemble des tags dans le même cluster, tandis que les tags ambigus sont ceux appartenant à l'intersection des clusters. Dans la dernière étape, nous avons amélioré l'algorithme de Heymann et al. [Heymann, 2006] à l'aide d'une nouvelle mesure de généralité appelé FDU (fréquence par utilisateurs distincts) pour extraire la hiérarchie des tags.

Ces approches mesurent la similarité entre les tags en se basant sur la ressource indépendamment de l'annotateur, ou sur l'utilisateur, quelle que soit la ressource, et la plupart d'entre elles ne traite pas le problème d'ambiguïté ou n'en donne pas une solution formelle. En outre, la majorité de ces approches n'explicitent pas les relations hiérarchiques entre les tags.

3.3.2 Approches basées sur les règles d'association

Cette classe d'approches applique les règles d'association dans le processus d'extraction d'ontologie pour la découverte de connaissances qui sont déjà implicitement présentes. Schmitz et al. [Schmitz 2006] proposent une approche qui adapte la structure tridimensionnelle des folksonomies pour qu'elle soit exploitable dans la recherche des règles d'association s'effectuant généralement sur des tables bidimensionnelles. Ils présentent tout d'abord un aperçu systématique de projection d'une folksonomie sur une structure à deux dimensions. Puis ils montrent les résultats de recherche des règles de la projection sélectionnée sur le

système del.icio.us [23](un système collaboratif de tagging qui permet aux utilisateurs d'annoter, de gérer et de partager des pages Web).

Jäschke et al. [Jäschke 2008], étendent la tâche du datamining pour la découverte des itemsets fermés fréquents à des structures de données tridimensionnelles. Cela fait situer leur approche sur la confluence des domaines de recherche des règles d'association et l'analyse des concepts formels FCA [Lehmann 1995] où chaque triplet est constitué d'un ensemble d'utilisateurs, un ensemble de tags et un ensemble de ressources.

Ces approches génèrent une représentation hiérarchique des tags, mais les relations entre les tags dans les différents niveaux hiérarchiques ne sont pas définies sémantiquement, en outre il n'y a pas de stratégie pour traiter les tags ambigus.

3.3.3 Les approches basées sur la sémantique

Ces approches visent à associer des entités sémantiques aux tags comme un moyen de définir formellement leur signification.

Angeletou et al. ont proposé une approche automatique afin d'enrichir les tags de la folksonomie avec la sémantique formelle en les associant à des concepts définis dans des ontologies en ligne [Angeletou, 2008]., L'ensemble du processus proposé est représenté à la figure 3.1.

Cantador et al. présentent une approche automatique pour associer des tags à des concepts d'ontologie de domaine à l'aide de Wikipedia [24] comme une représentation intermédiaire partagée entre les tags et les classes de l'ontologie [Cantador 2008].

Garcia-Silva et al. ont proposé une approche pour associer des tags à des ressources DBpedia [Auer 2007] au moyen de la sélection de la page Wikipedia

23. https ://delicious.com/
24. http ://en.wikipedia.org/

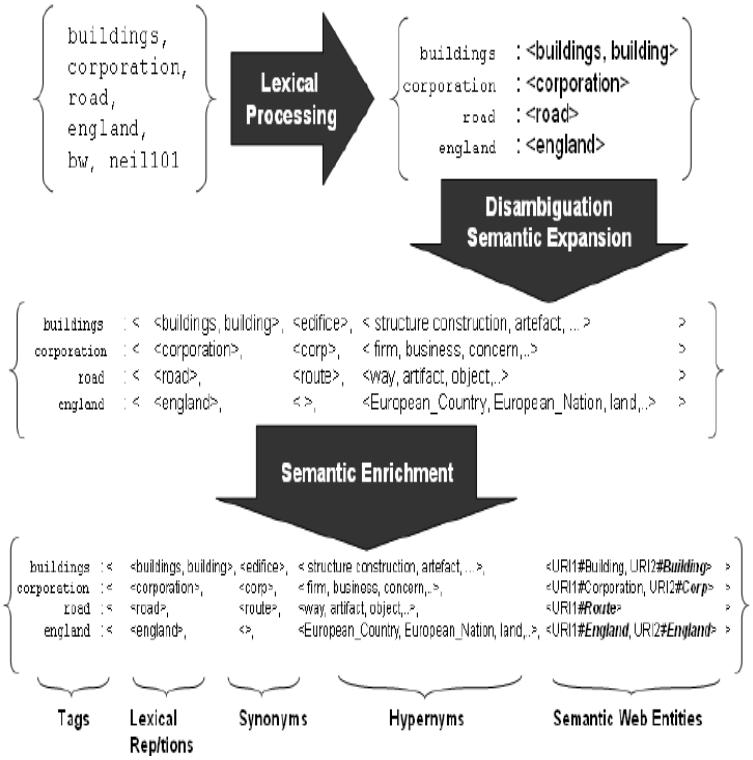

FIGURE 3.1 – L'approche de [Angeletou, 2008]

qui représente mieux le tag destiné, ce qui signifie le contexte où le tag a été utilisé [Garcia-Silva 2009].

Cette classe d'approches a besoin d'une ontologie de haut niveau existante comme une structure de base. L'absence d'ontologies qui correspondent bien aux tags des folksonomies est l'un des principaux obstacles à l'application de ces approches. Par exemple, WordNet est largement utilisé, cependant, les méthodes fortement basées sur WordNet obtiennent souvent des résultats médiocres parce qu'un bon nombre des tags n'existe pas dans WordNet.

3.3.4 Les approches hybrides

Dans cette section, nous présentons quelques approches intégrant des techniques multiples.

Giannakidou et al. ont proposé une approche statistique pour découvrir la sémantique des tags [Giannakidou 2008]. Cette approche est basée sur une mesure

de similarité composée de la cooccurrence des tags et de la similarité sémantique entre eux extraite de ressources externes (ontologies, thésaurus, etc.), ensuite l'approche regroupe les tags dans des groupes disjoints. Cela signifie qu'un tag peut appartenir à un seul groupe et donc si un tag a plusieurs sens différent, l'approche identifiera uniquement le sens le plus fréquent de ce tag.

Specia et al. proposent une approche semi-automatique d'extraction de la sémantique des tags qui emploie des techniques de clustering basées sur la cooccurrence entre les tags et des techniques pour aligner ces clusters à des éléments des ontologies (concepts, propriétés, instances, etc..) [Specia, 2007]. Toutefois, cette approche n'explicite pas les tags ambigus, ce qui nécessite d'analyser chaque cluster pour les retrouver. En outre, l'activité d'identification de la sémantique dans cette approche exige une analyse manuelle des ontologies provenant d'un moteur de recherche Web sémantique comme Swoogle [25]. Le processus global est illustré dans la figure 3.2

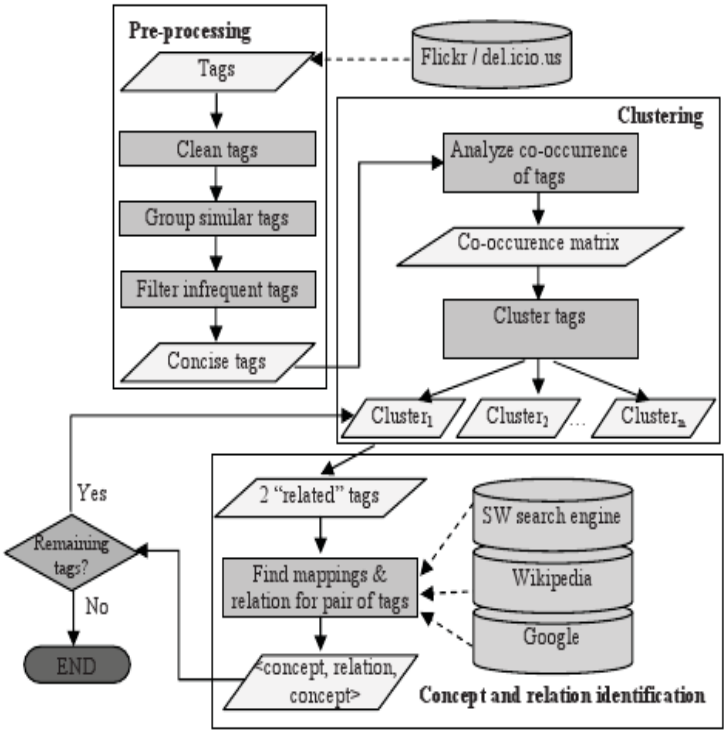

FIGURE 3.2 – l'approche de [Specia, 2007].

25. http ://swoogle.umbc.edu/

Lin et al. proposent une approche pour extraire la structure ontologique des folksonomies qui exploite la puissance des règles d'association et d'ontologie de haut niveau comme WordNet [Lin 2009](figure 3.3), mais cette ontologie bien qu'elle améliore la qualité des résultats, elle est considérée en même temps comme une limitation de cette approche du fait qu'une ontologie qui correspond mieux aux folksonomies n'est pas facile à retrouver.

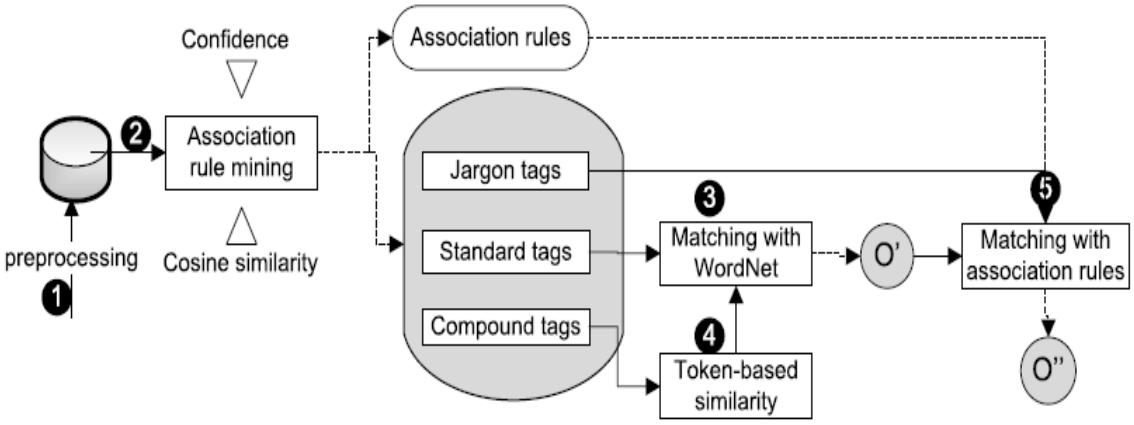

FIGURE 3.3 – l'approche de [Lin 2009].

Dans la section suivante, nous détaillons ces approches en les projetant dans un processus général de génération de la sémantique à partir des folksonomies. Cela dans le but de pouvoir comparer ces approches entre elles et surtout pour pouvoir délimiter les principaux inconvénients dans cet axe de recherche, afin de mieux mettre en valeur les qualités de notre approche par la suite.

3.4 Processus général de génération d'ontologies à partir des folksonomies

Dans cette section, nous présentons un processus unifié qui permet de comprendre, évaluer et classer les différentes approches pour l'association de la sémantique aux tags. Ce processus se compose d'un ensemble d'étapes en commun

retrouvée dans la plupart des propositions analysées. Le processus général est représenté dans la Figure 3.4

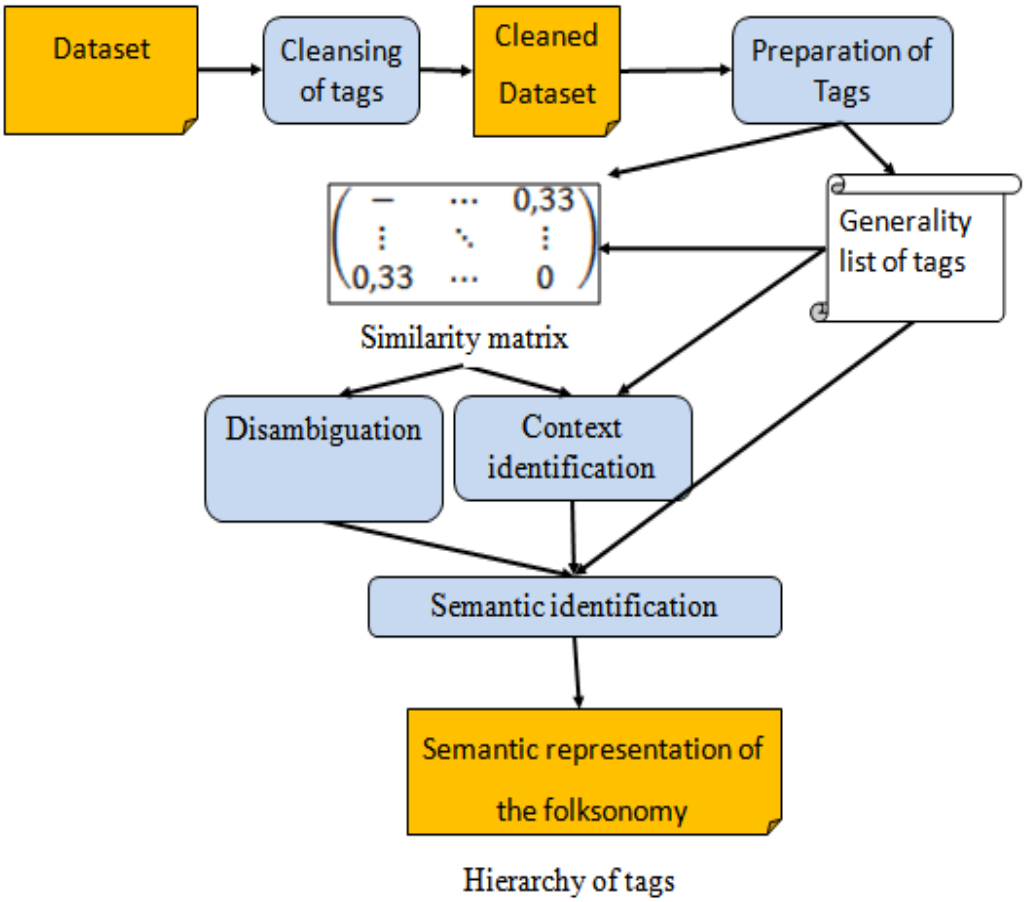

FIGURE 3.4 – Processus général pour l'extraction de la sémantique à partir des folksonomies.

La plupart des approches commence par définir les sources de données et certaines d'entre elles décrivent explicitement comment elles recueillent des informations provenant de ces sources de données. Par exemple, certaines équipes de recherche mettent au point des programmes spécialisés d'analyse de folksonomies lorsque les API qui font cette activité ne sont pas disponibles ou disponibles mais avec couverture limitée. Pour notre analyse, les détails de la façon d'obtenir les données ne sont pas si importantes. Ce qui nous intéresse sont les prétraitements qu'ils mettent en oeuvre pour sélectionner et nettoyer les données, si elles existent.

3.4.1 Nettoyage de données

Les informations de tagging peuvent contenir des tags bruyants et des incompatibilités. Lorsque les tags sont introduits par un mécanisme de marquage non-contrôlé, les utilisateurs font souvent des erreurs grammaticales (par exemple Barclona à la place de Barcelone), ils utilisent les concepts de tags indistinctement au singulier, pluriel ou des dérivations linguistiques (blog, blogs), parfois ils ajoutent des adjectifs, adverbes, prépositions ou pronoms au concept principal du tag (belle voiture, à lire), ou bien ils emploient des synonymes et acronymes qui pourraient être convertis en un seul tag (biscuit et cookies, ny et New York).
En outre, les mécanismes d'encodage et de stockage des tags utilisés par les systèmes sociaux modifient souvent les tags mis en place par les utilisateurs : ils peuvent transformer les espaces blancs (San Francisco, sanfrancisco) et des caractères spéciaux dans les tags (Los Angeles pour Los Ángeles , Zurich à la place de Zürich), etc.

Ainsi, alors qu'il est possible de recueillir des informations à partir de plusieurs sites de folksonomie, tels que Flickr ou del.icio.us, l'incohérence serait source de confusion et de perte d'informations lorsque les informations de tagging sont comparées. Par exemple, si un utilisateur a annoté des photos d'un récent séjour à New York avec « NYC », mais aussi des pages pertinentes de del.icio.us avec « NewYork » , la corrélation sera perdue.

Afin de faciliter l'analyse des données et l'intégration de folksonomie, les tags doivent être filtrés et mis en correspondance avec un vocabulaire commun. Ainsi, la première activité identifiée est appelée **Nettoyage de données**. Cette activité comprend des filtres qui tiennent en compte la fréquence, les caractéristiques lexicales, les caractéristiques morphologiques, ou même la langue des tags. En ce qui concerne la fréquence du tag, il est mesuré selon le nombre de fois où le tag a été utilisé pour annoter une ressource, ou le nombre de fois où le tag a été utilisé par différents utilisateurs.

L'activité qui suit le nettoyage de données est la préparation des tags.

3.4.2 Préparation des tags

Cette étape consiste à agréger les informations de tagging qui forment les trois modes des folksonomies selon deux modes, calculer les similarités entre les tags, et dans certaines approches calculer le degré de généralité des tags composant la folksonomie.

3.4.2.1 Arégations des informations de tagging

Dans la figure 3.5 nous avons un exemple d'une petite folksonomie où deux utilisateurs annotent trois ressources par trois tags. Chaque lien de la folksonomie est constitué de trois parties : un utilisateur associe un tag à une seule ressource. L'idée est de projeter ces liens tripartites de la folksonomie en représentations bipartites en agrégeant les données selon un contexte donné. Si nous voulons avoir les similarités entre les tags, tois contextes existent :

1. le contexte tag-tag, où l'on considère la cooccurrence des tags dans les messages.
2. le contexte de tag-ressource, où l'on considère les associations des tags via les ressources sur lesquelles ils sont utilisés.
3. le contexte utilisateur -tag, où l'on considère les associations des tags par les utilisateurs qui les utilisent.

Dans ce qui suit, nous allons examiner les différentes méthodes d'agrégation utilisées à savoir la projection proposée par Mika [Mika, 2007], l'agrégation distributionnelle présentée par Cattuto et al. [Cattuto, 2008], la macro-agrégation et l'agrégation collaborative proposée, plus tard, par Markines et al. [Markines, 2009]).

L'agrégation par projection. Ce type d'agrégation a été tout d'abord étudié par [Mika, 2007] et consiste à projeter l'hypergraphe tripartite d'une folksonomie sur différents types de graphiques à deux modes correspondant à chacun des contextes décrits ci-dessus.

L'hypergraphe d'une folksonomie est donnée par $H(F) = \langle V, E \rangle$ avec l'ensemble des sommets $V = U \bigcup T \bigcup R$. Et l'ensemble des arêtes $E = u,t,r|(u,t,r) \in F$

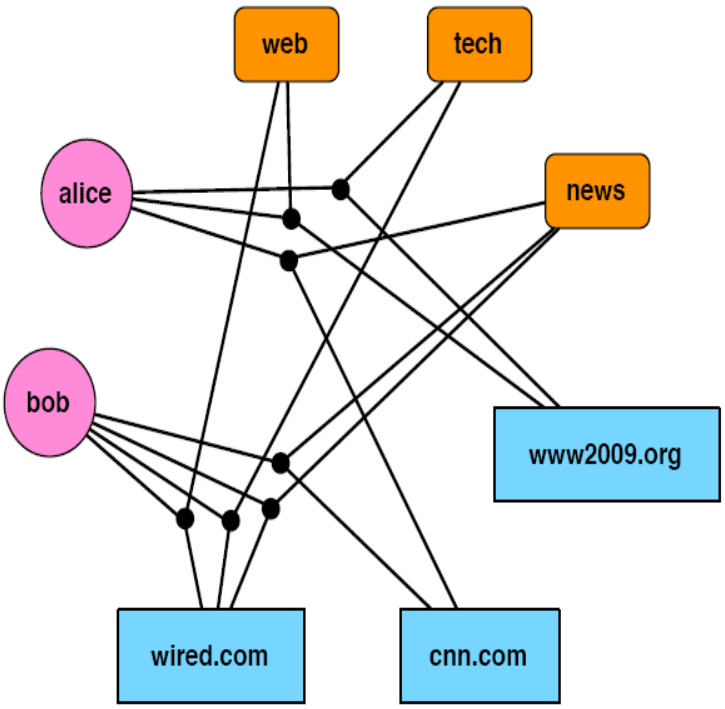

FIGURE 3.5 – L'exemple de folksonomie proposé par Markines et al [Markines, 2009].

où R est l'ensemble des ressources, U l'ensemble des utilisateurs, et T le jeu de tags.

L'étude de Mika a porté sur les contextes tag-ressource et tag-utilisateur, mais plus généralement, l'agrégation par projection consiste à créer une matrice d'affiliation des tags avec un des autres éléments du contexte envisagé.

Ainsi, dans le cadre du contexte Tag-tag, cette affiliation correspond à la cooccurrence des tags. Cette matrice sera constituée de $card(\mathrm{T})$ lignes et $card(\mathrm{T})$ colonnes, et chaque cellule (i, j) a la valeur 1 si le tag t_i apparait au moins une fois avec le tag t_j. Sinon la cellule a la valeur 0.

De même, la matrice d'affiliation dans le contexte tag-ressource représente l'affiliation entre chaque tag et chaque ressource, et entre chaque tag et chaque utilisateur dans le contexte utilisateur-tag. Dans le tableau 3.1, nous donnons l'exemple de l'agrégation par projection pour le contexte tag-ressource pour la

folksonomie exemple donnée à la figure 3.5.

	cnn.com	www2009.org	wired.com
news	1	0	1
web	0	1	1
tech	0	1	1

TABLE 3.1 – Exemple d'agrégation par projection dans le contexte Tag-ressource correspondant à l'exemple de folksonomie de Markines et al. [Markines, 2009] donné dans la figure 3.5

Mika extrait ensuite, depuis la projection de Tag-ressource un graphe pondéré représentant les connexions entre les tags basés sur les ressources et à partir de la projection Tag-utilisateur un autre graphe ayant les usagers comme relations entre tags. Dans le cas des associations basées sur les utilisateurs, pour une paire donnée de tags, les poids du graphe sont donnés par le nombre d'utilisateurs qui ont utilisé les deux tags et normalisés par le nombre d'utilisateurs, correspondant ainsi à la similarité de Jaccard entre le Tags (le détail de cette mesure de similarité est donné ci-dessous).

La figure 3.6 montre un exemple de ce dernier type de graphiques. Ce graphique a été construit à partir d'un extrait de delicious.com reliant deux tags lorsque le poids du lien entre eux est supérieur à un seuil donné.

Agrégation distributionnelle Des méthodes plus élaborées d'agrégation des données du taggin font usage de l'hypothèse distributionnelle qui considère que les mots utilisés dans des contextes similaires ont tendance à être sémantiquement connexes [Firth, 1957]. Cette hypothèse a été exploitée par Cimiano [Cimiano, 2006] à des fins d'extraction d'ontologies à partir de textes. Cette façon de regrouper les données de tagging considère chaque élément du contexte d'agrégation comme une dimension dans un vecteur représentant les autres éléments. Par exemple si l'on considère la représentation vectorielle des tags dans le contexte tag-ressources, chaque dimension correspond à l'une des ressources de la folksonomie.

L'Agrégation distributionnelle consiste donc à calculer les composants du vecteur v_t qui représente chaque tag t pour chaque contexte comme suit :

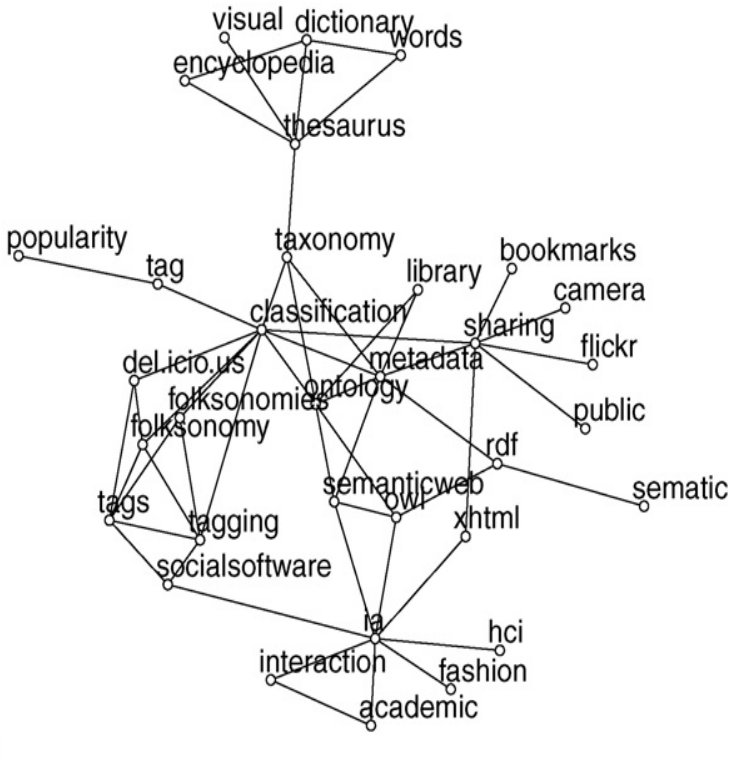

FIGURE 3.6 – Des tags del.icio.us liés grâce à une projection de la folksonomie sur le contexte Tag-utilisateur

1. Le contexte de tag-tag : chaque entrée du vecteur du tag t v_t correspond à la co-occurrence du tag t avec toutes les autres tags, à l'exception du tag t avec lui-même où une valeur de 0 est donnée. Il s'agit d'éviter de considérer deux tags similaires simplement lorsqu'ils se produisent ensemble, mais plutôt quand ils ont des profils de co-occurrence similaires, c'est-à-dire lorsqu'ils se produisent avec les mêmes autres tags.
2. Le contexte Tag-ressource : pour un tag t, le vecteur v_t est construit en comptant combien de fois un tag $t$est utilisé pour annoter une certaine ressource r.
3. Le contexte Tag-utilisateur. Pour un tag t le vecteur v_t est construit en comptant combien de fois un tag t est utilisé par un certain utilisateur u.

Si nous choisissons le contexte tag-ressource, la représentation de la matrice correspondant à l'agrégation distribustionnelle pour la folksonomie exemple donné à la figure 3.5 sera ressemblée à ce que nous donnons dans le tableau 3.2. Par exemple, le vecteur des tag « news » dans le contexte Tag-ressource sera $v_{news} =$

$(2, 0, 1)$.

	cnn.com	www2009.org	wired.com
news	2	0	1
web	0	1	1
tech	0	1	1

TABLE 3.2 – Exemple d'une agrégation distributionnelle dans le contexte tag-ressources de l'exemple de folksonomie de al Markines et al. [Markines, 2009] donné dans la figure 3.5

Macro agrégation et agrégation collaborative La projection et l'agrégations distributionnelle sont considérées par Markines et al. [Markines, 2009].comme non incrémentale, étant donné que la matrice de similarité entière doit être recalculée une fois que chaque utilisateur ajoute une nouvelle annotation. Ainsi, ces types d'agrégation peuvent ne pas être scalable, puisque leur temps de calcul n'augmente pas constamment avec la croissance de la folksonomie. Pour surmonter cette limitation, Markines et al. proposent un autre type d'agrégation, appelée "macro-agrégation » qui consiste à :

1. considérer l'agrégation et calculer la la similarité correspondante de chaque utilisateur séparément.
2. agréger entre tous les utilisateurs, c'est-à-dire, calculer la somme des similarités locale pour l'ensemble de données de chaque utilisateur.

Cela dit, quand un utilisateur u fournit une nouvelle annotation, Il n'est pas nécessaire de recalculer la similarité globale de la folksonomie entière , mais uniquement pour cet utilisateur.

En outre, et afin de tenir compte de la similarité entre deux ressources annotées par les mêmes utilisateurs mais avec aucun tag en commun Markines et al. ont proposé une autre façon de calculer les similarités locales, appelées « agrégation collaborative ».

L'objectif de la méthode d'agrégation collaborative est réalisé par l'ajout d'un vecteur spécial "utilisateur tag" (respectivement " utilisateur ressource ") à toutes les ressources (respectivement tags) de l'utilisateur u. Prenons l'exemple des tags « news » et « web » l'utilisateur « alice », de la folksonomie de figure 3.5.

Si l'on ajoute la ressource virtuelle « aliceR » à la matrice binaire représentant Alice (tableau 3.3), nous allons avoir une similarité locale non nulle entre les tags « news » et « web » pour l'utilisateur « alice » puisque ces deux tags apparaissent ensemble sur la ressource virtuelle « aliceR ». Ensuite, la mesure de similarité est calculé comme dans le cas de macro-agrégation en additionnant les similarités locales entre utilisateurs.

	cnn.com	www2009.org	wired.com	aliceR
news	1	0	0	1
web	0	1	0	1

TABLE 3.3 – Représentation de matrice binaire pour la méthode d'agrégation collaborative pour les tags « news » et « web » pour l'utilisateur « Alice ». La dernière colonne est la ressource virtuelle ajoutée

3.4.2.2 Mesures de similarité.

Différentes mesures de similarité peuvent être appliquées sur les données de deux modes résultant des méthodes d'agrégation que nous avons décrit ci-dessus (projection, distributionnelle, macro et les agrégations collaboratives). Markines et al. ont appliqué et évalué six types de mesures de similarité qui peuvent être effectuées : matching, overlap, coefficient dice, jacquard, cosine, et la similarité information mutuelle.

Nous allons présenté brièvement ici la similarité de Jaccard utilisée par Mika[Mika, 2007] (mais seulement dans le cas de l'agrégation de projection), la similarité cosinus utilisée par Cattuto et al. [Cattuto, 2008] mais seulement dans le cas d'agrégation distributionnelle, et la mesure de similarité information mutuelle présentée et évaluées par Markines et al. [Markines, 2009] qui a donné les meilleurs résultats par rapport aux autres mesures.

Prenons deux tags x_1 et x_2, avec X_1 et X_2 leurs représentations vectorielles. Chaque entrée de ces vecteurs est représentée par $w_{1,xy}$ et $w_{2,xy}$ (pour x_1 et x_2 respectivement) où y correspond au contexte de l'agrégation. Par exemple, dans le contexte tag-ressource, y représente les différentes ressources, chacune prise comme une dimension du tag x. Pour les agrégations par projection, le vecteur

binaire X peut être vu comme un ensemble, et $y \in X$ signifie que $w_{xy} = 1et|X| = \Sigma_y w_{xy}$.

Dans l'agrégation distributionnelle, chaque élément w_{xy} d'un vecteur X correspond à la valeur de la dimension y. Par exemple, dans le contexte tag-ressource, w_{xy} correspond au nombre de fois que le tag x est utilisé sur la ressource y.

De même, dans la macro agrégation et l'agrégation collaborative, pour un seul utilisateur u, $y \in X_u$ est équivalant à $w_{u,xy} = 1et|Xu| = \sum_y w_{u,xy}$.

Similarité de Jaccard. La similarité de Jaccard est une mesure de similarité entre deux représentations vectorielles, elle est définie pour l'agrégation par projection comme :

$$\sigma(x_1, x_2) = \frac{|x_1 \bigcap x_2|}{|x_1 \bigcup x_2|}$$

Et dans l'agrégation distributionnelle :

$$\sigma(x_1, x_2) = \frac{\sum_{y \in \{x_1 \cap x_2\}} Logp(y)}{\sum_{y \in \{x_1 \cup x_2\}} Logp(y)}$$

où p(y) est donnée par N(x,y)/N(y) où N(x,y) pour le contexte tag-ressource (resp.Tag-tag) est le nombre de fois que x est utilisé pour annoter la ressource y (resp. le nombre de fois que les tags x et y sont utilisés les deux ensemble). N(y) est le nombre total des ressources (resp. le nombre total de tags).

Dans la macro agrégation et les agrégations collaboratives, nous considérons la similarité pour chaque utilisateur u, et l'expression évolue un peu plus :

$$\sigma(x_1, x_2) = \frac{\sum_{y \in \{x_1^u \cap x_2^u\}} Logp(y/u)}{\sum_{y \in \{x_1^u \cup x_2^u\}} Logp(y/x)}$$

où p(y/u) est la valeur locale de p(y) pour utilisateur u, c'est-à-dire $p(y/u) = N(x,y)_u/N(y)_u$ où $N(x,y)_u$ dans le contexte tag-ressources (resp. tag-tag) représente le nombre de fois où x est utilisé pour annoter la ressources y par l'utilisateur u (resp. le nombre de fois le tag x et le tag y apparaissent ensemble et $N(y)_u$ est le nombre total de ressources annotées par l'utilisateur u(resp. le nombre total de tags utilisés par l'utilisateur u).

Similarité cosinus La similarité cosinus entre le tag x_1 et le tag x_2 est donnée par la valeur de la distance cosinus entre les représentations de deux vecteurs X_1 et X_2. Pour l'agrégation par projection cette similarité est calculée par l'équation

$$\sigma(x_1,x_2) = \frac{|X_1 \bigcap X_2|}{\sqrt{X_1.X_2}}$$

Pour l'agrégation distributionnelle, elle s'écrit :

$$\sigma(x_1,x_2) = \frac{X_1.X_2}{\sqrt{\|X_1\|_2.\|X_2\|_2}}$$

Et dans la macro agrégation et l'agrégation collaborative, le calcul est basé sur les valeurs locales pour chaque utilisateur u :

$$\sigma(x_1,x_2) = \frac{|x_1^u \bigcap x_2^u|}{\sqrt{|x_1^u|.|x_2^u|}}$$

la mesure de similarité information mutuelle. Markines et al. [Markines, 2009] ont proposé une nouvelle mesure de similarité appelée information mutuelle.
La similarité information mutuelle $\sigma(x_1,x_2)$ entre les deux tags x_1 et x_2 est définie pour l'agrégation par projection et l'agrégation distributionnelle comme :

$$\sigma(x_1,x_2) = \sum_{y_1 \in X_1} \sum_{y_2 \in X_2} p(y_1,y_2) log \frac{p(y_1,y_2)}{p(y_1)p(y_2)}$$

Où, dans l'agrégation par projection, p(y) est la fraction de tags annotant la ressource y, et la probabilité p(y_1,y_2) est la fraction des tags annotant les deux ressources $y_1 et y_2$ donnée par $p(y_1,y_2) = \frac{\sum_x w_{xy1} w_{xy2}}{N(x)}$ où $\sum_x w_{xy1} w_{xy2}$ calcule le nombre de tags qu'ont annoté les deux ressources $y_1 et y_2$, et N(x) correspond au nombre total des tags. Pour l'agrégation distributionnelle la normalisation pour p(y) et $p(y_1,y_2)$ se fait à travers la matrice entière plutôt que sur les colonnes c'est à dire $p(y) = \frac{\sum_x w_{xy}}{\sum_{r,t} w_{rt}}$ et $p(y_1,y_2) = \frac{\sum_x min(w_{xy1},w_{xy2})}{\sum_{r,t} w_{rt}}$ avec $\sum_{r,t} w_{rt}$ est la somme de toutes les entrées de la matrice.

Dans le cas de la macro agrégation et de l'agrégation collaborative, l'information mutuelle locale pour un utilisateur u est donnée par :

$$\sigma(x_1,x_2) = \sum_{y_1 \in X_1^u} \sum_{y_2 \in X_2^u} p(y_1,y_2/u) log \frac{p(y_1,y_2/u)}{p(y_1/u)p(y_2/u)}$$

où les probabilités locales p(y/u) sont données par $p(y/u) = \frac{N(u,y)}{(N(u)+1)}$ où N(u, y) est le nombre de tags utilisés par u pour annoter la ressource y, Tandis que N(u) est le nombre total des tags de u. Les probabilités sont normalisées de la même façon à l'agrégation par projection, mais dans ce cas pour la représentation binaire de chaque utilisateur. La valeur globale de cette similarité est obtenue en additionnant dans l'ensemble de tous les utilisateurs ces similarités locales.

3.4.2.3 Mesures de généralité :

Comprendre les différents niveaux de généralité des tags (ou abstraction des tags) est essentiel pour identifier les relations hiérarchiques entre les concepts. Cette tâche est pertinente pour toutes les applications qui en bénéficient pour une meilleure compréhension de la sémantique des tags, par exemple les algorithmes d'extraction des ontologies ou de clustering, les systèmes de recommandations de tag ou les systèmes de navigation des folksonomies. Benz et al. [Benz, 2011] ont proposé et évalué un ensemble de mesures de généralité en se basant sur plusieurs intuitions (mesures basées sur la fréquence, sur l'entropie des tag et les mesures basées sur le degré de centralité des tags dans le graphe de cooccurrence) que nous présentons dans cette section.

Mesure basée sur la fréquence. Ces mesures se basent sur l'idée que les tags les plus abstraits sont les tags les plus utilisés pour annoter les ressources web. C'est dire, le degré de généralité d'un tag t est donné par sa fréquence d'utilisation. Formellement :

freq(t) = card $\{(u, t', r) \in Y : t = t'\}$.

Mesure basée sur l'entropie. Une autre intuition découle de la théorie de l'information : l'entropie mesure le degré d'incertitude associé à une variable aléatoire. En considérant l'application des tags comme un processus aléatoire, nous pouvons attendre que les tags les plus généraux montrent une distribution plus uniforme, parce qu'ils sont probablement utilisés à un niveau relativement constant pour annoter un large éventail de ressources. Par conséquent, les tags

les plus abstraits auront une entropie plus élevée. Cette approche a également été utilisée par Heymann et al. [Heymann, 2008] pour capturer la généralité des tags dans le contexte de la recommandation des tags. formellement cette mesure est définie par :

$$entr(t) = \sum_{t' \in cooc(t)} p(t'|t) log p(t'|t).$$

où cooc(t) est l'ensemble des tags qui ont apparu ensemble avec t, et $p(t'|t) = \frac{w(t',t)}{\sum_{t'' \in cooc(t)} w(t'',t)}$. Avec w(t',t) étant le poids de cooccurrences .

Mesure de centralité. Dans la théorie des graphes la centralité d'un nœud $v \in V$ du réseau G est généralement une indication de l'importance de ce nœud [Wasserman, 1994]. En appliquant cette notion à notre problème, la centralité peut également être envisagée comme le niveau d'abstraction des tags en suivant l'intuition que les termes les plus abstraits sont également les plus importants. Benz et al. [Benz, 2010] ont adopté trois degrés de centralité (degré de centralité, la proximité et la centralité d'intermédiarité). Chaque mesure peut être appliquée à un graphe de tags G :

1. le degré de centralité calcule tout simplement le nombre des voisins directs d(v) pour un nœud v dans un graph G(V,E)
   $dc(v) = \frac{d(v)}{|V|-1}$
2. Selon la centralité d'intermédiarité, un noeud a une centralité élevée s'il peut être trouvé dans plusieurs plus courts chemins calculés pour des paires de noeuds différentes
   bc(v)=$\sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$
   σ_{st} désigne le nombre de chemins les plus courts entre s et t, $\sigma_{st}(v)$ est le nombre de chemins les plus courts entre s et t en passant par v. Comme son calcul est évidemment très couteux, elle est souvent approximée [Brandes, 2007] en calculant les chemins les plus courts entre quelques points seulement.
3. Finalement selon la centralité de proximité, nous considérons un tag plus central plus son chemin le plus court vers les autres noeux est court.
   $cc(v) = \frac{1}{\sum_{t \in V}}$

$d_G(v,t)$ où dG(v,t) est le chemin le plus court entre v et t.

3.4.3 Identification de contexte

Par contexte, on entend l'ensemble des tags, que nous tenons compte pour comprendre la signification du tag. Cela aidera à identifier les groupes de tags associés ou d'associer un concept sémantique qui défini formellement la signification du tag. Le contexte en linguistique est défini comme les mots voisins d'un mot. Cette notion de contexte peut être appliquée pour les tags d'une folksonomie de deux façons :

la première c'est d'utiliser les tags qui figurent ensemble pour annoter la même ressource ou un ensemble de ressources, la seconde consiste à utiliser des tags qui sont utilisés conjointement par le même utilisateur ou groupe d'utilisateurs.

En outre la notion de contexte peut inclure des informations linguistiques, telles que les synonymes ou les hyperonymes ou autres variations morphologiques des tags.

En revanche, le contexte peut inclure également des métadonnées de tag ou de ressource, et aussi des informations telles que les coordonnées de l'emplacement et les horodatages [Kennedy, 2007].

Comme mentionné précédemment, l'un des principaux problèmes avec les folksonomies est que la plupart des tags ont plus qu'un seul sens. Cette ambiguïté fournit des résultats inexacts et non pertinents lorsque ces tags sont utilisés pour rechercher et récupérer des informations.

Par conséquent, une étapes de désambigüisation est une étape importante dans le processus d'association de la sémantique. L'activité de désambigüisation est réalisée généralement à l'aide de ressources sémantiques extérieures, comme WordNet et toutes les informations de contexte de tag.

Certaines approches, telles que celle présentée par Au Yeung et al[Au Yeung, 2009] et Specia et al. [Specia, 2007], utilisent des techniques de clustering afin de regrouper les tags selon les ressources qu'ils annotent ou aux utilisateurs qui annotent en les utilisant ces tags. Dans les deux cas, selon ces méthodes, si un tag est

utilisé pour annoter des groupes de ressources différents ou si un tag est utilisé par différents groupes d'utilisateurs, alors le tag est considéré ambigü. En général, ce type d'analyse se concentre plus sur l'identification de l'existence d'une ambiguïté, plutôt que sur l'identification de la vraie signification d'un tag.

3.4.4 Identification de la sémantique

Enfin, la dernière activité est l'identification de la sémantique dans laquelle la sémantique formelle et explicite du tag est extraite.

Cette activité comporte la correspondance entre les tags et les entités sémantiques, ou l'identification des relations entre les tags ou entre les entités sémantiques. La correspondance entre les tags et les entités sémantiques comme des classes ou des instances est réalisée à l'aide d'ontologies prédéfinies ou ontologies récupérées pendant l'exécution au moyen des moteurs de recherche du Web sémantique. Cette correspondance peut entraîner plusieurs entités sémantiques pour un tag [Angeletou, 2008], d'où l'agrégation de ces entités est nécessaire afin de déterminer lequel d'entre eux, représente le même topic et qui ne le fait pas. Dans le cas où les entités sémantiques sont associées à plus d'une rubrique, une tâche de désambigüisation doit être effectuée. En outre, cette activité pourrait utiliser des techniques de clustering pour identifier les groupes de synonymes ou des mesures de réseau social, comme le coefficient de clustering et la centralité locale, pour identifier les groupes de termes plus génériques et plus spécifiques semblables aux relations trouvées dans un dictionnaire des synonymes [Mika, 2007].

3.5 Comparaison des approches

Dans cette section, nous présentons une comparaison entre les approches qui extraient des relations sémantiques entre les tags, nous nous basant dans cette comparaison sur le processus général que nous avons présenté dans la section précédente. le tableau 3.4 résume cette comparaison.

3.5.1 Nettoyage des données

Dans [Mika, 2007], la tâche de nettoyage de cette approche est limitée au filtrage des tags de la folksonomie ayant moins de dix éléments classés en vertu de

Approche	Type	Source de données	Séléction et nettoyage des données	Disambiguisation	Identification de la sémantique
Heymann et al. 2006	clustering	Delicious et CiteULike	NON	Non	Oui
Mika, 2007	clustering	Delicious	Oui	Non	Oui
Hamasaki et al, 2007	clustering	Delicious	NON	Oui	Non
Giannakidou et al, 2008	Hybride	Flikcr	Oui	Non	Non
Jäschke et al, 2008	Règles d'association	Delicious et Bibsonomy	Oui	Non	oui
Specia et Motta, 2007	Hybride	Delicious et Flikcr	Oui	Oui	Oui
Angeletou et al, 2008	sémantique	Flikcr	Oui	Oui	Oui
Cantador et al, 2008	sémantique	Delicious et Flikcr	Oui	Non	Oui
Garcia Silva et al, 2009	sémantique	Flikcr	Oui	Oui	Oui
Benz et al, 2010	clustering	Delicious	Oui	Oui	Oui
Schmitz et al, 2006	Règles d'association	Delicious	Oui	Non	Oui
Lin et al, 2009	hybride	Flickr et CiteULike	Oui	Non	Oui
Marouf et al, 2013	clustering	Delicious et Flikcr	Oui	Oui	Oui

TABLE 3.4 – Tableau récapitulatif des approches d'extraction de la sémantique à partir des folksonomies

ceux-ci et les personnes qui ont utilisé moins de cinq tags. Le dataset de Delicious utilisé pour tester l'approche contient 51.852 anotations, 30.790 ressources, 10.198 utilisateurs et 29.476 tags distincts.

Aucune règle générale pour la sélection de données n'est donnée dans [Hamasaki, 2007].

Heymann et al. [Heymann, 2006] n'effectuent aucun prétraitement sur les données qu'ils choisissent pour leurs expérimentations, les sources de données qu'ils utilisent sont Delicious et CiteULike. Par contre l'amélioration qu'apportent Benz et al. [Benz, 2010] sur leur travail consiste à améliorer la qualité des informations du tagging choisies par supprimer les tags utilisés seulement une fois par un seul utilisateur. Cela dans l'objectif d'avoir trop de valeurs nulles de cooccurrence.

Dans la première version de notre approche [Marouf, 2013], nous avons suivi les mêmes étapes de [Cantador 2008] et [Garcia-Silva 2009], mais sur une folksonomie différente. La folksonomie que nous avons choisie pour nos expérimentation est Flickr. Au début les données contenaient $|U| = 319,686$ utilisateurs, $|R| = 28,153,045$ ressources, et $|T| = 1,607,879$ tags. Après le nettoyage nous avons obtenu 21,925 images, 2000 termes et 1962 utilisateurs.

Schmitz et al. [Schmitz 2006] n'ont réalisé aucun prétraitement sur les données collectées. La folksonomie choisie est celle de Delicious avec $|U| = 75,242$ utilisateurs, $|T| = 533,191$ tags et $|R| = 3,158,297$ ressources.

Dans [Jäschke 2008], les données sélectionnées pour les expérimentation sont un snapshot de la folksonomie choisie. Par exemple, dans le cas de Delicious, les auteurs utilisent toutes les informations de marquage introduites dans le système avant le 16 juin 2004. Cet ensemble de données contenait plus de 3,3 K utilisateurs, environ 30,5 K de Tags, plus de 220 K de ressources, avec près de 617 K de liens. En ce qui concerne Bibsonomy, les auteurs ont choisi toutes les données jusqu'au 23 novembre 2006, à l'exclusion de toute insertion automatique (par exemple les publications DBLP) ainsi que tout tag par défaut (par exemple Imported). Les données de Bibsonomy contiennentt presque 45K annotations, 262 utilisateurs et plus 11 K ressources (publications) avec près de 6 K tags distincts.

Angeletou et al. [Angeletou, 2008], ont utilisé une folksonomie de Flickr contenant 250 photos choisies aléatoirement avec 2819 tags individuels. Après l'élimination des tags non anglais et les tags contenant des chiffres et des caractères spéciaux, les données qui leur restent sont 226 photos avec 1146 tags. Ensuite ils génèrent toutes les représentations lexicales pour chaque tag pour résoudre les incompatibilités entre les différentes conventions de nommage utilisées dans les

folksonomies, ontologies et thésaurus tels que WordNet.

Dans [Cantador 2008], le jeu de tags initial contient 28.550 tags, recueillis de Delicious et de Flickr. L'activité de sélection de données et de nettoyage, représentée à la Figure 3.7, met l'accent sur le filtrage des tags courts (comme des tags avec une seule lettre) ou trop longs (tags avec plus de 25 caractères). En outre, les caractères spéciaux sont convertis à leur forme de base, et les tags avec une faible fréquence ou qui sont des mots vides sont supprimés. Ensuite, chaque tag est recherché dans WordNet. S'il n'existe pas, il est recherché dans le mécanisme *Google did you mean* qui sert à corriger n'importe quelle faute d'orthographe possible ou de fractionner tout tag composé (tags composés des mots concaténés). Sinon, les tags sont censés être des acronymes, sigles ou noms propres. Dans ce dernier cas, ces tags sont recherchés dans Wikipedia pour une représentation convenue. En outre, les tags morphologiquement semblables sont regroupés dans un tag unique à l'aide d'un algorithme de stemming. Enfin, les tags qui sont des synonymes non ambigus sont fusionnés.

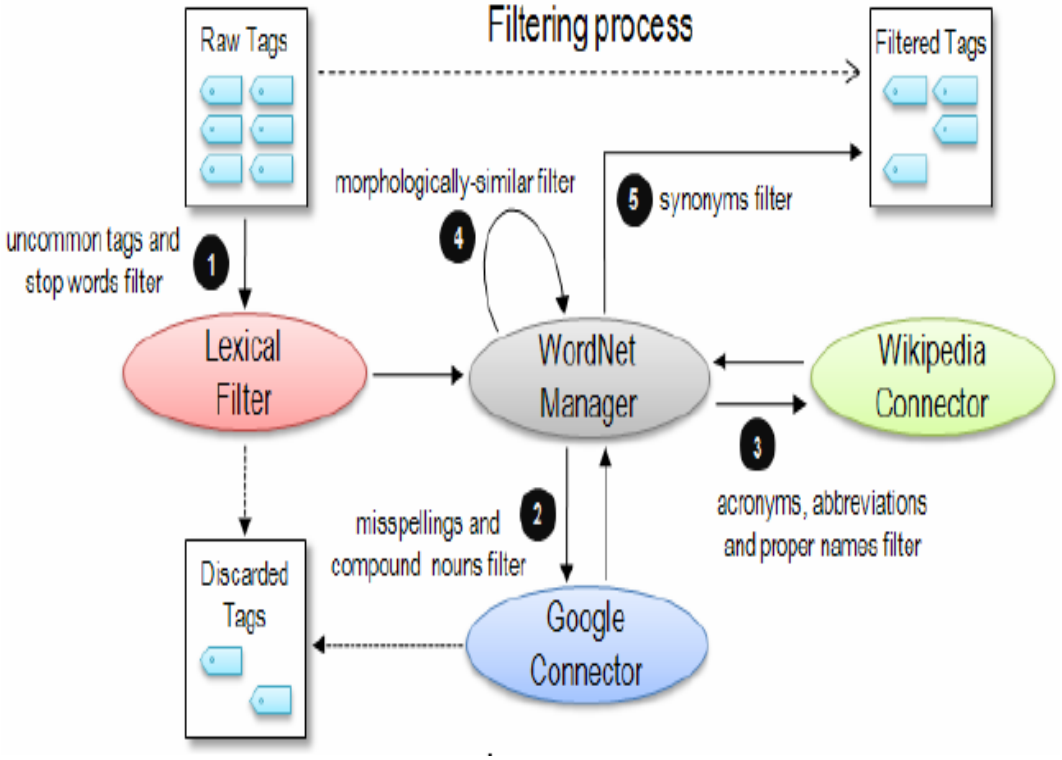

FIGURE 3.7 – Les étapes de nettoyage d'après [Cantador 2008].

Dans [Garcia-Silva 2009], la sélection de données et l'activité de nettoyage comprennent les mêmes étapes présentées par Cantador et al, puisque les deux approches partagent la même dataset pour tester leurs approches. Cependant, les auteurs ne fournissent pas des statistiques concernant le jeu de données utilisé pour tester l'approche.

Dans [Giannakidou 2008], l'approche effectue la normalisation orthographique où les différentes variations orthographiques d'un tag sont mappées à une version normalisée du tag. Les tags peu fréquents sont filtrés avec ceux qui n'ont pas un concept correspondant dans WordNet, qui est la ressource terminologique que les auteurs utilisent pour le calcul de similarité sémantique. L'ensemble de données utilisé pour tester l'approche contient 3000 ressources. De ces tags les 30 tags les plus fréquents sont extraits pour être analysés.

Specia et al. [Specia, 2007] ont utilisé les folksonomies de Delicious et Flickr, leur prétraitement consiste à garder seulement les tags qui commencent par une lettre, ensuite regrouper les tags similaires en utilisant la similarité de Levenshtein avec un grand seuil pour fusionner les tags avec des variations morphologique mineures.

Lin et al. [Lin 2009] ont effectué leurs expérimentations sur deux collections de données de Flickr et CiteULike. Le prétraitement consiste à supprimer les ressources avec seulement un tag et les tags non anglais tout en évitant de supprimer les mots de jargon et les mots composés.
Le tableau 3.5 récapitule les techniques utilisées pour sélectionner et nettoyer les tags.

3.5.2 Préparation des données

La comparaison entre les techniques d'agrégation et les mesures de similarités est représentée dans le tableau 3.6

3.5.3 Identification du contexte

Dans [Mika, 2007], l'activité d'identification de contexte est effectuée de manière distincte pour chacun des ontologies résultant. Dans le cas de l'ontologie O_{ci}

Approche	prétraitement
Mika, 2007	filtrage des tags de la folksonomie ayant moins de dix éléments classés en vertu de ceux-ci et les personnes qui ont utilisé moins de cinq tags
Hamasaki et al.	-
Heymann et all	-
Cantador et al., 2008	Filtrage des tags courts ou trop longs des tags avec une faible fréquence et les mots vides, les caractères spéciaux sont convertis à leur forme de base, recherche dans WordNet, Google did you mean et Wikipedia, algorithme de stemming
Marouf et al.,2013	Même étapes que Cantador et al
Schmitz et al., 2006	-
Jäschke et al., 2008	Exclusion de toute insertion automatique (par exemple les publications DBLP) ainsi que tout tag par défaut (par exemple Imported)
Angeletou et al., 2008	Elimination des tags non anglais et les tags contenant des chiffres et des caractères spéciaux, génèrent toutes les représentations lexicales pour chaque tag
Garcia-Silva et al.,2009	Même étapes que Cantador et al
Giannakidou et al., 2008	Normalisation orthographique où les différentes variations orthographiques d'un tag sont mappées à une version normalisée. Les tags peu fréquents sont filtrés avec ceux qui n'ont pas un concept correspondant dans WordNet
Specia et al, 2007	Garder seulement les tags qui commencent par une lettre, ensuite regrouper les tags similaires en utilisant la similarité de Levenshtein .
Lin et al, 2009	Supprimer les ressources avec seulement un tag et les tags non anglais tout en évitant de supprimer les mots de jargon et les mots composés.

TABLE 3.5 – Les techniques utilisées pour sélectionner et nettoyer les tags

le contexte est défini comme les tags qui figurent avec un tag particulier lorsqu'ils sont utilisés pour annoter une ressource. Dans le cas de l'ontologie O_{ac} le contexte est défini comme les tags qui apparaissent avec un tag particulier lorsqu'ils sont utilisés par un même utilisateur ou par un groupe d'utilisateurs.

Approche	**Représentation de la folksonomie**	**Contexte**	**Mesure de similarité**	**Mesure de généralité**
Mika., 2007	Agrégation par projection	Tag-utilisateur et Tag-resource	Jaccard	Jugement de la communauté des utilisateurs
Hamasaki et al.,2007	Agrégation par projection + informations des voisins	Tag-utilisateur et Tag-resource	cooccurrence	-
Heymann et all	Agrégation distributionnelle	Tag-resource	Cosine	Degré de centralité
Benz et all.,2010	Agrégation distributionnelle	Tag-resource	Cosine	Degré de centralité
Marouf et al.,2013	Macro agrégation	Tag-resource	CDU + Cosine	FDU
Schmitz et al., 2006	projection d une folksonomie sur une structure à deux dimensions	-	Probabilité conditionnelle	-
Jäschke et al., 2008	Un ensemble de triconcepts selon les principes de FCA [Lehmann 1995]	-	Probabilité conditionnelle	-
Cantador et al.,2008	-	-	Appariement morphologique	-
Angeletou et al., 2008	-	-	Cooccurrence + similarité de Wu and Palmer [Wu 1994]	-
Garcia-Silva et al.,2009	-	-	Cooccurrence + Cosine	-
Giannakidou et al., 2008	Agrégation par projection	Tag-resource	Social Similarity SOS + Semantic Similarity SeS	-
Specia et al, 2007	-	-	Cooccurrence + Cosine	-
Lin et al, 2009	Agrégation par projection	Tag-resource -	Probabilité conditionnelle + Cosine	-

TABLE 3.6 – La préparation des données dans les différentes approches

Dans le travail de [Hamasaki, 2007], le contexte de tag forme les tags de l'utilisateur ainsi que les tags utilisés par ses voisins. Les informations de tagging des voisins pourraient aider à surmonter l'absence d'informations de tagging pour un utilisateur particulier.

Le travail de Heymann et al. [Heymann, 2006] ne s'intéresse pas à l'identification du contexte, il extrait directement la hiérarchie des tags.

Benz et al. [Benz, 2010] considère le contexte d'un tag, les 10 tags les plus fréquents parmi les tags cooccurents avec ce tag.

Dans [Marouf, 2013], nous avons effectué un clustering flou pour regrouper les tags similaires dans un même cluster. Dans ce travail, le contexte d'un tag est constitué des tags appartenant au même cluster auquel il appartient.

L'identification du contexte de l'approche de [Jäschke 2008] consiste d'extraire tous les triconcepts fréquents sur les informations sélectionnées afin d'obtenir un ensemble de triplets, où chaque triplet contient un ensemble d'utilisateurs, un ensemble de tags et un ensemble de ressources. Chaque utilisateur dans l'ensemble des utilisateurs annote chaque ressource de l'ensemble des ressources avec touts les tags dans l'ensemble des tags.

Dans [Angeletou, 2008], pour chaque tag, sont générés toutes ses représentations lexicales possibles, comme le singulier, pluriel ou les divers types délimité par des tags composés. Le contexte de tags est défini comme l'ensemble des tags filtrés ainsi que leurs représentations lexicales.

Dans[Cantador 2008], l'activité d'identification de contexte récupère les informations de Wikipedia pour chaque tag, y compris l'URL de la page Wikipedia et la liste de ses catégories.

L'approche de [Garcia-Silva 2009] essaie de lever l'ambiguïté sur la signification de chaque tag dans chaque poste utilisateur. Par conséquent, le contexte de tag est défini comme les tags qui figurent ensemble dans le poste de l'utilisateur.

Giannakidou et ses collègues, [Giannakidou 2008] définissent le contexte d'un tag comme l'ensemble des tags cités conjointement avec le tag donné pendant l'annotation. En outre, le contexte d'une ressource est défini comme les tags que les utilisateurs ont assignée.

En ce qui concerne l'approche de [Specia, 2007], l'objectif de l'activité d'identification de contexte ici, c'est de construire des clusters de tags connexes tout d'abord, le contexte d'un tag est défini comme l'ensemble des tags utilisés conjointement avec le tag courant pour annoter une ressource, ou quand ils sont utilisés par le même utilisateur. Pour représenter le contexte d'un tag les auteurs utilisent un vecteur dont le nombre d'éléments est égal au nombre de tags distincts dans la folksonomie, et les valeurs de chaque entrées correspondent au nombre de fois que le tag est utilisé conjointement avec le tag qui correspond à la position actuelle. Dans le cas où l'élément du vecteur correspond au tag qui consiste à identifier le vecteur, la valeur de cet élément est la fréquence d'utilisation de ce tag dans la folksonomie. Ensuite, chaque tag est comparé avec les autres tags à l'aide de leurs vecteurs de contexte afin de trouver des tags similaires. Le tableau 3.7 établit une comparaison entre les approches selon leurs méthodes d'identification de contexte.

3.5.4 Désambigüisation

Contrairement aux approches de [Mika, 2007], [Jäschke 2008], [Schmitz 2006] [Cantador 2008], [Heymann, 2006][Giannakidou 2008] et [Lin 2009] qui ne traitent pas explicitement le problème d'ambiguïté, l'approche de [Hamasaki, 2007] propose un algorithme pour la désambiguïsation. L'algorithme est basé sur l'idée que si un tag est utilisé pour annoter différentes ressources par différents groupes d'utilisateurs (voisins), le tag peut avoir des significations différentes. Ce qui signifie que dans le cas contraire, le tag a un seul sens (ou des sens très semblables). L'algorithme proposé traite chaque tag d'utilisateur comme un pré-concept, et puis ces pré-concepts sont fusionnés s'ils ont les mêmes étiquettes et partagent les mêmes utilisateurs/ressources ou utilisateurs voisins.

Benz et al. [Benz, 2010] réalisent une classification hiérarchique basée sur le lien moyen sur les représentations vectorielles selon le contexte de chaque tag pour extraire les tags ambigus et pour définir leurs sens .

Dans notre travail initial [Marouf, 2013], nous avons utilisé les même résultats de clustering qui ont défini le contexte des tags, cette fois pour detecter et désambiguiser les tags ambigus. Ces derniers représentent les tags qui se trouvent dans l'intersection de plusieurs clusters.

Approche	Identification de contexte
Mika, 2007	Tags cooccurrents
Hamasaki et al., 2007	Tags cooccurrents
Heymann et al. 2006	-
Marouf et al., 2013	Tags appartenant au même cluster
Schmitz et al., 2006	-
Jäschke et al., 2008	Les tags d'un même treillis de concepts
Angeletou et al., 2008	Tags cooccurrents
Cantador et al., 2008	-
Garcia-Silva et al, 2009	Tags cooccurrents
Giannakidou et al., 2008	Tags cooccurrents
Specia et al, 2007	Tags appartenant au même cluster
Benz et al. 2010	10 tags cooccurrents les plus fréquents
Lin et al, 2009	-

TABLE 3.7 – Comparaison des approches selon leurs méthodes d'identification de contexte

Dans [Angeletou, 2008], les tags et leurs contextes sont pris comme entrée dans l'activité de désambigüisation. Dans cette activité, si un tag a plusieurs sens dans WordNet, alors la hiérarchie de ses sens, tels que figurant dans WordNet est utilisée pour calculer la similarité avec tous les tags dans l'ensemble des tags, et ainsi désambigüiser le tag. La similarité entre les sens est calculée à l'aide de la mesure de similarité de Wu et Palmer [Wu 1994].

L'activité de désambigüisation dans l'approche de [Garcia-Silva 2009] commence par récupérer un ensemble de pages candidates de Wikipedia liées au tag ambigu en utilisant le répertoire de sens Tagora (TSR) [26]. En outre, pour chaque page Wikipedia les termes les plus fréquents sont extraits. Ainsi, le tag et son

26. http ://tagora.ecs.soton.ac.uk

contexte peuvent être comparés contre chacune des pages candidates de Wikipedia en mesurant le chevauchement des termes dans le contexte avec les termes fréquents de chaque page Wikipedia. Le tag, avec son contexte, ainsi que les pages Wikipedia sont représentés sous forme de vecteurs qui sont ensuite comparés au moyen de la mesure de similarité cosine. Le vecteur de page Wikipedia le plus similaire au tag et son contexte est sélectionné comme le sens plus probable de ce tag.

Dans [Specia, 2007], lorsqu'un tag est ambigu, il peut avoir plus d'un modèles de cooccurrence. Ainsi, les tags similaires trouvés dans l'identification du contexte peuvent inclure des tags avec des significations différentes. L'activité de désambigüisation analyse chaque groupe de tags similaires afin de trouver des clusters de tags associés basés sur la cooccurrence.

Le tableau 3.8 résume ces approches en vue de la désambigüisation.

3.5.5 Identification de la sémantique

Dans l'approche de [Mika, 2007], l'activité d'identification sémantique est également effectuée différemment pour chacune des ontologies que génère cette approche. Un graphique des concepts est construit pour O_{ci} où les arrêtes indiquent que les deux concepts (tags) ont été utilisés ensemble pour annoter une ou plusieurs ressources. Ces liens sont pondérés par le nombre de ressources annotées à l'aide de ces deux tags. De même, un graphe des concepts est également construit pour O_{ac} où les arrêtes indiquent que les deux concepts ont été utilisés ensemble par un ou plusieurs utilisateurs. Ces liens sont pondérés par le nombre de personnes qui ont utilisé les deux tags. Après la génération de graphique, une tâche de clustering est réalisée afin d'identifier les mots spécifiques à l'intérieur de chaque cluster et les termes généraux. Les principes de la théorie des ensembles sont ensuite appliqués pour définir les relations entre les concepts comme plus générique et plus spécifique. La sortie est représentée par deux hiérarchies de concepts Oac et Oci.

Hamasaki et al. [Hamasaki, 2007] ont composé les deux matrices tag-ressource et tag-utilisateur de Mika et al. [Mika, 2007] avec une troisième matrice qui repré-

Approche	Méthode de désambigüisation.
Hamasaki et al, 2007	le tag utilisé pour annoter différentes ressources par différents groupes d'utilisateurs est considéré ambigu.
Benz et al., 2010	classification hiérarchique basée sur le lien moyen sur les représentations vectorielles selon le contexte de chaque tag pour extraire les tags ambigus et pour définir leurs sens.
Marouf et al., 2013	Les tags appartenant à l'intersection des clusters sont considérés ambigus, et leurs sens sont à partir des autres tags appartenant à ses clusters.
Angeletou et al., 2008	Utilisation de WordNet pour déterminer les tags ambigus et calcul de similarité entre les différents sens et les autres tags de la folksonomies.
Garcia-Silva et al, 2009	Désambigüisation par récupération d'un ensemble de pages candidates de Wikipedia liées au tag ambigu.
Specia et Motta., 2007	Un tag est ambigu s'il a plus d'un modèles de cooccurrence, pour le désambigüiser il faut analyse chaque groupe de tags similaires afin de trouver des clusters de tags associés basés sur la cooccurrence.

TABLE 3.8 – Comparaison des approches selon leurs méthodes de désambigüisation

sente les liens entre les utilisateurs pour générer la sémantique de la folksonomie sous forme d'un réseau de relations entre les tags.

Heymann et al. [Heymann, 2006] ont généré une hiérarchie de tags à partir de la folksonomie à l'aide de leur algorithme proposé.

Benz et al. [Benz, 2010] ont amélioré la hiérarchie de Heymann et al. pour qu'elle prenne en compte les tags ambigus.

Dans [Marouf, 2013] nous avons à notre tour amélioré le processus de génération de la hiérarchie de Benz et al. en utilisant FDU comme mesure de généralité au lieu du degré de centralité et cela grâce à sa simplicité de calcul et de ses résultats satisfaisants.

L'approche de Schmitz et al. [Schmitz 2006] donnent comme résultats un ensemble de règles d'association entre les tags de la forme $a \rightarrow b$ qui peuvent être considérées comme des règles de subsumption entre les tags (a subsume b) et ainsi être utilisées pour générer des structures taxonomiques.

Dans [Jäschke 2008], l'identification de la sémantique des tags est effectuée en sélectionnant dans les triplets trouvés dans les activités antérieures l'ensemble de tags qui intéressent l'étude. Puis, pour chaque ensemble de tags un treillis de concept est créé. Un treillis de concept est un regroupement conceptuel hiérarchique des tags. Le contexte formel du treillis de concept se compose des ressources annotées par un certain nombre d'utilisateurs en utilisant au moins un tag de l'ensemble des tags. Le treillis est composé également de tags qui ont été utilisés par la majorité des utilisateurs pour annoter une ressource. La représentation graphique de ce réseau de concept pourrait alors servir aux ingénieurs d'ontologie pour construire manuellement une hiérarchie de concept qui correspond à la folksonomie originale.

L'identification de la sémantique de l'approche de [Angeletou, 2008] met l'accent sur la correspondance entre l'ensemble de tags aux entités ontologiques en utilisant le moteur de recherche Watson [27]. Pour chaque tag, plusieurs entités ontologiques peuvent être trouvées, et sont alors intégrées afin de regrouper les entités ontologiques similaires.

Au cours de l'étape d'identification de la sémantique de l'approche de [Cantador 2008], les auteurs donnent à chaque tag un URI de concept en utilisant les informations de contexte des tags, qui contient le nom de la page Wikipedia et la catégorie Wikipedia précédemment associées à chaque tag. À cette fin, les termes dans le contexte sont comparés avec les classes d'ontologie de domaine et les classes d'ontologies. Enfin, une instance de chacune des classes d'ontologie est créée. Les URI de ces instances sont les noms de la page Wikipedia, et les catégories sont assignées en tant qu'étiquettes des instances.

Dans [Garcia-Silva 2009], l'identification sémantique est réalisée en sélectionnant le concept de DBpedia correspondant à la page Wikipedia sélectionnée

27. http ://watson.kmi.open.ac.uk/

lors de la phase précédente. Dans DBpedia chaque concept à un *foaf* : page object property qui relie le concept de DBpedia avec la page correspondante de Wikipédia. En utilisant cette propriété le concept de DBpedia pourra être facilement identifiée depuis l'URL de la page de Wikipedia.

Dans[Giannakidou 2008], l'activité d'identification sémantique crée un graphe où les ressources, et les tags les plus fréquents dans la folksonomie, sont représentés comme des sommets. Les arêtes du graphe associent les ressources aux tags. Une arête entre une ressource et un tag existe si la valeur de leur similarité est au-dessus d'un certain seuil. Dans cette approche, chaque ressource est représentée par l'ensemble des tags utilisés pour l'annoter, de ce fait la similarité entre un tag et une ressource est calculée comme la valeur maximale de similarité du tag avec chacun des tags utilisés pour annoter la ressource. La similarité entre deux tags est la somme de leurs similarités sociales pondérées par leurs similarités sémantiques. La similarité sociale repose sur la cooccurrence des deux tags pendant l'annotation des ressources. Pour la similarité sémantique les auteurs proposent de mapper les tags aux concepts dans une ressource sémantique. Ensuite, la similarité sémantique est calculée proportionnellement par la distance entre les concepts dans la ressource sémantique. Le graphe biparti relatif des tags et des ressources est alors partitionné à l'aide d'un algorithme de clustering spectral dont le but de créer des clusters disjoints afin que les éléments de même cluster aient une similarité élevée et les éléments dans des clusters différents aient une similarité faible.

Dans l'approche de [Specia, 2007] pour chaque cluster de tags l'activité d'identification sémantique est effectuées manuellement. Un utilisateur utilise un moteur de recherche Web sémantique (par exemple, Swoogle).Il cherche des ontologies contenant les paires de tags dans le cluster. Si une ontologie trouvée contient le paire de tags, alors les informations sur les tags (type, parents, domaine) sont utilisées pour établir des relations entre eux.
Dans le tableau 3.9 nous donnons une comparaison entre ces approches concernant la génération de la sémantique des tags.

Approche	Identification de la sémantique.
Mika, 2007	Deux hiérarchie de concepts Oac et Oci.
Hamasaki et al., 2007	Réseau qui représente de relations entre les tags.
Heymann et al., 2006	Hiérarchie de tags.
Benz et al., 2010	Hiérarchie de tags.
Marouf et al., 2013	Hiérarchie de tags.
Schmitz et al., 2006	Ensembles de règles d'association entre les tags.
Jäschke et al., 2008	Création des treillis de concept.
Angeletou et al., 2008	La correspondance entre l'ensemble de tags à des entités ontologiques en utilisant le moteur de recherche Watson.
Cantador et al., 2008	Donner à chaque tag un URI de concept qui contient le nom de la page Wikipedia et la catégorie Wikipedia.
Garcia-Silva et al, 2009	Sélection de concept DBpedia correspondant à la page Wikipedia sélectionnée.
Giannakidou et al., 2008	Création d'un graphe où les ressources, et les tags les plus fréquents dans la folksonomie, sont représentés comme des sommets.
Specia et Motta., 2007	Les informations sur les tags (type, parents, domaine) dans l'ontologie qui contient ses tags sont utilisées pour établir des relations entre eux.

TABLE 3.9 – Comparaison entre les approches selon leurs méthodes d'identification de la sémantique

3.6 Conclusion

Dans ce chapitre, nous avons présenté plusieurs approches pour extraire les relations sémantiques entre les tags en analysant les structures des folksonomies. Nous avons présenté les principales approches connexes. ensuite, nous avons projeter ces travaux sur un processus général de génération de la sémantique à partir des folksonomie. Dans la section nous avons présenté les différentes mesures de similarité et de généralité qui peuvent être utilisées pour trouver des relations de subsumption entre tags, ou regrouper les tags similaire dans des clusters. Dans le tableau 3.6, nous avons rapporté les différents types de mesure de similarité utilisés par les approches de l'étude.

Dans une étude comparative, Cattuto et al. [Cattuto, 2008] et plus tard Koerner et al. [Koerner, 2010] ont réalisé que la mesure cosinus calculée dans

l'agrégation distributionnelle dans le cadre du contexte tag-tag a donné une sémantique de bonne qualité à un coût de calcul raisonnable. Dans le chapitre suivant nous allons proposé une nouvelle mesure de généralité et de similarité basée sur une nouvelle agrégation, ensuite nous évaluerons les deux mesures par rapport à la mesure de généralité *fréquence* et les mesure de similarité cosine et cooccurrence, cela, parce que ces mesures prouvent leur efficacité.

Chapitre 4

Une nouvelle approche pour l'extraction de la sémantique à partir des folksonomies

4.1 Introduction

Après avoir introduit dans les chapitres précédents les caractéristiques du-Web2.0 et des folksonomies ainsi que différentes pistes relatives à une complémentarité entre les folksonomies et les ontologies, nous allons présenter ici notre approche dans ce contexte et nous montrons comment nos propositions comblent les limites des approches existantes. Le travail présenté dans cette thèse repose sur une combinaison de plusieurs techniques telles que le clustering, la désambigüisation et la génération de la hiérarchie. Notre approche diffère des autres approches par l'emploi de nouvelles mesures de similarité et de généralité, et par l'utilisation du clustering flou au lieu du clustering dur (hard ou crisp-clustering) pour définir le contexte des tags et pour déterminer les différents sens des tags ambigus. Dans cette section, nous décrivons notre approche qui vise à extraire les structures ontologiques à partir des folksonomies en effectuant les étapes qui suivrons. Le procédé global est illustré sur la figure 4.1.

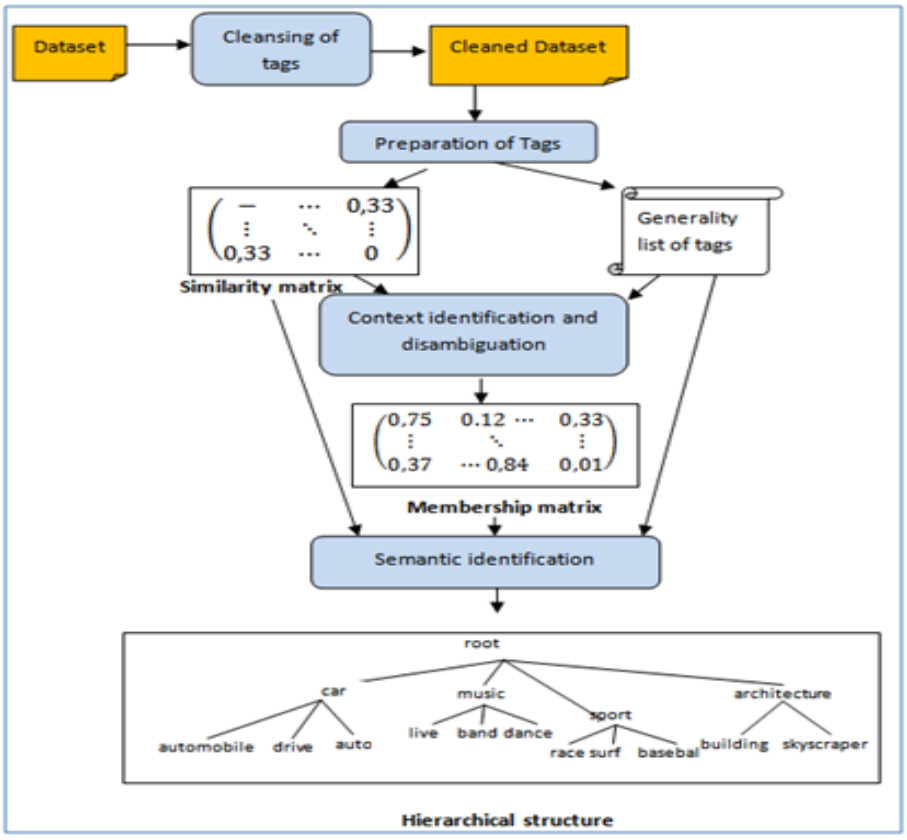

FIGURE 4.1 – Architecture de l'approche proposée.

4.2 Nettoyage des tags

Comme les utilisateurs des folksonomie ont la liberté de choisir n'importe quel mot-clé pour catégoriser leur contenu, ils appliquent leurs propres règles d'orthographe et de marquage (par exemple les noms singuliers ou pluriels, les verbes conjugués). Par conséquent, les tags sont pollués et doivent être nettoyés. de ce fait et avant d'entamer l'analyse des données de la folksonomie, nous devons nettoyer le dataset. Pour ce faire, on procède comme décrit dans l'algorithme 1. Nous commençons par supprimer tous les tags ayant une fréquence d'apparition réduite, des mots vides, et les tags qui n'ont aucun sens (lignes 3,4) (ce qui signifie qu'une étape de calcul des fréquences des tags est effectuée avant l'exécution de l'algorithme). Les mots vides sont éliminés sur la base d'une comparaison des tags avec des mots d'un fichier contenant des mots vide anglais tels que les pronoms, les prépositions et articles. Ensuite, l'algorithme vérifie si le tag a une signification ou non en vérifiant son existence dans le dictionnaire Wordnet (ligne 6) ; cette étape est réalisée à l'aide de l'API Java JWNL ; si le tag existe dans Wordnet, il est conservé et transmis à l'étape suivante pour qu'il soit transformé en son radical ou racine (ligne 9). Sinon, le tag n'est pas pris en considération et ne sera pas présent dans les étapes à venir de l'approche (ligne 7). Dans cette thèse nous traitons uniquement les tags en anglais pour avoir une structure hiérarchique cohérente. Une solution multilingue est de générer autant des structures hiérarchiques que les langues utilisées dans la folksonomie.

4.3 Préparation des tags :

Dans cette étape, nous générons une liste des tags dans leur ordre descendant de la généralité et une représentation vectorielle des tags. Le degré de généralité est basé sur la mesure FDU qui calcule la fréquence d'utilisation d'un tag donné par différents utilisateurs. La représentation vectorielle est basée sur une nouvelle mesure de similarité que nous appellons NCDU (Normalised Co-occurrences in Distinct Users). Ces deux mesures vont être utilisées dans les étapes qui suivent.

Algorithm 1 Nettoyage des tags

```
Require: datasetFile ; {un fichier de données représentant les postes dela folk-
    sonomie}
 1: while not end of (datasetFile) do
 2:   datasetFile.readLine() ;
 3:   if tag is in(stop words, meaningless tags, infrequent tags) then
 4:     DatasetFile.getNextline ;
 5:   else
 6:     Exist = Wordnet.check(tag) ;{look for the tag in WordNet}
 7:     if not Exist then
 8:       DatasetFile.getNextline ;
 9:     else
10:       tag =Stemming(tag) ; {The stemming task reduces tags to their stem
          or root}
11:       CleanDataSetFile.add(Tag) ;
12:     end if
13:   end if
14: end while.
```

Avant de décrire les mesures de FDU (Frequency by Distinct Users) et NCDU, nous donnons d'abord une définition formelle de la folksonomie et ses éléments.

Definition 4.3.1 (Folksonomie)**.** *Une folksonomie est définie comme un tuple* $F = \{T, U, R, Y\}$*, où* T *est l'ensemble des tags qui composent le vocabulaire exprimé par la folksonomie ; U, R sont respectivement les ensembles d'utilisateurs et de ressources qui annotent et sont annotées par les tags de* T*; et* $Y = \{u,t,r\} \in U \times T \times R$ *est l'ensemble des affectations (annotations) de chaque tag à une ressource par un utilisateur u.*

Definition 4.3.2 (Poste)**.** *Un poste est un triplet* $(u, t_{ur}, r) avec u \in U$*,* $r \in R$ *et l'ensemble non vide* $t_{ur} = \{t \in T \setminus (u,r,t) \in Y\}$*.*

La figure 4.2 montre un exemple d'une folksonomie. Dans ce qui suit, nous utilisons cet exemple pour illustrer notre approche.

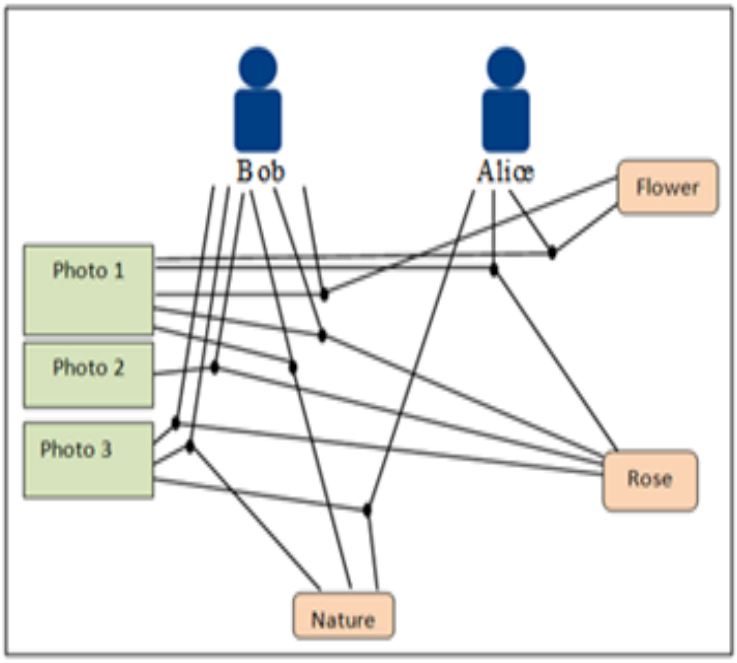

FIGURE 4.2 – Un exemple d'une folksonomie.

4.3.1 Agrégation des informations de tagging

Comme les traitements de calcul de similarité et de généralité ne sont pas bien développés pour les ensembles de données de trois modes tels que les folksonomies, nous devons réduire les triples par un. Diverses solutions ont été proposées pour calculer les similarités entre les tags et entre les ressources selon deux modes de données [Markines, 2009]. Contrairement au travail de Markines et all[Markines, 2009] qui utilisent un contexte tag-ressource pour modéliser la folksonomie et par la suite agréger à tous les utilisateurs, nous utilisons une représentation binaire tag-tag pour chaque utilisateur au lieu de la représentation tag-ressource, nous regroupons ensuite toutes les représentations pour avoir une représentation binaire globale des similarités entre les tags de la folksonomie. Cette représentation améliore les performances d'exécution, car le nombre de ressources est largement supérieur par rapport au nombre des tags (surtout après la tâche de nettoyage), de sorte que la représentation tag-ressource nécessite plus d'espace mémoire et de temps de calcul. Cette nouvelle agrégation comble les lacunes des agrégations par projection et distribution, du fait qu'elle est bien incrémentale et scalable. Cela revient au fait qu'il n'est pas nécessaire de recalculer la

matrice globale de la folksonomie à chaque ajout d'une nouvelle annotation, mais seulement pour l'utilisateur qui l'a fournit. En outre, pour chaque utilisateur, elle calcule directement la matrice tag-tag sans passer par la matrice tag-ressource, comme le font l'agrégation collaborative et la macro-agrégation. Ainsi, elle permet d'éviter beaucoup de calcules.

Les valeurs des représentations binaires tag-tag par utilisateur sont $w_u(t_1,t_2) \in \{0,1\}$, où t_1 et t_2 sont des paires de tags. $w_u(t_1,t_2) = 0$ signifie que t_1 et t_2 n'apparaissent pas ensemble dans un poste associé à l'utilisateur "u" . $w_u(t_1,t_2) = 1$ signifie que t_1 et t_2 se trouvent au moins une fois ensemble dans les postes relatifs au utilisateur "u" . Formellement : $\forall u \in U$, $(t_1,t_2) \in T$, $r \in R$

$$w_u(t_1,t_2) = \begin{cases} 1 & \text{si} \quad \exists t_{ur} \backslash (t_1,t_2) \in t_{ur} \\ 0 & \text{sinon} \end{cases} \tag{4.1}$$

Pour la folksonomie exemple de la Figure 4.2, la représentation binaire est montrée au tableau 4.1 et tableau 4.2 pour les utilisateurs Alice et Bob respectivement.

Tag	Flower	Rose	Nature
Flower	-	1	1
Rose	1	-	1
Nature	1	1	-

TABLE 4.1 – la matrice binaire de Bob

Tag	Flower	Rose	Nature
Flower	-	1	0
Rose	1	-	0
Nature	0	0	-

TABLE 4.2 – la matrice binaire de Alice

4.3.2 La mesure de généralité FDU

Une première intuition naturelle est que les tags les plus généraux sont tout simplement les plus souvent utilisés, car ils sont bien connus par les utilisateurs. Nous saisissons cette intuition dans la mesure de généralité "FDU".

Definition 4.3.3 (FDU). *La mesure de généralité FDU est une version adaptée de la mesure "fréquence" qui calcule le nombre d'utilisateurs distincts qui annotent des ressources avec un tag donné. Formellement, nous la définissons comme suit :*

$$\forall u \in U, t \in TFDU(t) = Card\{(u) \in U \backslash t \in t_{ur}\}. \tag{4.2}$$

Ci-dessous, nous présentons les valeurs de généralité des tags de l'exemple qui en résultent.

Tag	Flower	Rose	Nature
FDU	2	2	2

TABLE 4.3 – Les valeurs de FDU résultant

4.3.3 La mesure de similarité NCDU

NCDU est une version améliorée de CDU proposée dans [Marouf, 2013]. Elle est calculée en additionnant les matrices binaires des utilisateurs, puis en divisant sur le plus petit degré de généralité pour chaque paire de tags. Les avantage de cette nouvelle mesure ne se limitent pas seulement de prendre en compte les trois modes de la folksonomie tous à la fois, mais aussi qu'elle génère des valeurs de similarité normalisées et plus précises entre les tags.

Definition 4.3.4 (NCDU). *Formellement, nous définissons NCDU comme suit :*

$$\forall u \in U, t_1, t_2 \in TNCDU((t_1, t_2) = \frac{\sum_u w_u(t_1, t_2)}{min(FDU(t_1, t_2)} \tag{4.3}$$

Ci-dessous, nous rapportons la matrice résultante de la similarité pour l'exemple présenté précédemment

Tag	Flower	Rose	Nature
Flower	-	1	1/2
Rose	1	-	1/2
Nature	1/2	1/2	-

TABLE 4.4 – la matrice de similarité basée sur NCDU de l'exemple 4.2

Cette nouvelle mesure permet de comparer facilement les similarités entre les tags dans l'étape de génération de la hiérarchie. En outre, elle n'exige pas de

générer d'autres matrices de similarités telles que cosine ou autre, et cela grâce à la normalisation des valeurs qu'elle assure, ainsi que la flexibilité qu'offre notre nouvel algorithme de clustering de traiter les matrices creuses. Cet algorithme est au centre de la section suivante.

4.4 Identification de contexte et Désambiguïsation

Malgré les avantages des folksonomies, elles souffrent de divers problèmes ; l'ambiguïté (polysémie) des tags apparaît lorsque les utilisateurs annotent en utilisant le même tag dans différents domaines. De l'autre côté, le manque de contrôle de synonyme se produit lorsque différents tags sont utilisés pour le même concept. Plusieurs solutions sont proposées dans la littérature. Il y a des approches qui tentent d'identifier la signification réelle d'un tag en articulation avec les bases de connaissances structurés ([Angeletou, 2009]). Autres recherches appliquent des modèles probabilistes et des techniques de clustering sur l'espace des tags selon leur co-occurrence ([Weinberger, 2008], [Shepitsen, 2008], [Au Yeung, 2009]).

De même dans notre travail, nous suivons une stratégie de clustering, mais à la différence des approches précédentes, notre proposition offre les avantages suivants :

1. Au lieu d'utiliser des algorithmes de clustering standard, nous proposons d'appliquer une technique de clustering flou qui permet aux tags d'appartenir à plus d'un cluster. Les tags similaires appartiennent au même cluster, tandis que les tags ambigus se trouvent dans l'intersection de deux ou plusieurs clusters.
2. Au lieu d'utiliser la mesure de similarité de co-occurrence ou cosine, nous utilisons la nouvelle mesure de similarité NCDU.

La littérature fournit plusieurs exemples d'identification du contexte et de désambiguïsation. Ces approches considèrent les techniques de clustering durs comme un moyen d'assigner un ensemble de tags à des groupes disjoints. Les tags dans un même cluster sont plus semblables les uns aux autres qu'à ceux des autres groupes. Cette tâche de clustering détermine la signification des tags en

recueillant ses tags semblables, Mais des efforts supplémentaires doivent être fait pour aborder le problème d'ambiguïté. Dans [Specia, 2007], la désambiguïsation consiste d'analyser chaque groupe de tags similaires générés dans l'étape d'identification du contexte afin de trouver des tags avec des significations différentes. Benz et al [Benz, 2010] appliquent l'algorithme de clustering hiérarchique basé sur le lien moyen [Pantel, 2002] pour lever l'ambiguïté des tags.

Dans des travaux antérieurs et dans le but de la désambiguïsation des tags [Marouf, 2013], nous avons utilisé un algorithme de clustering flou appelé FCM (Fuzzy c-means) [Bezdek, 1981] pour regrouper les deux étapes d'identification de contexte et de la désambigüisation dans une même étape, et en même temps afin de faciliter la deuxième tâche. Le clustering flou offre la possibilité aux éléments de clustering de pouvoir appartenir à plusieurs clusters en même temps. En considérant les éléments d'un même cluster comme des éléments partageant un sens en commun, pour un tag, appartenir à plusieurs clusteurs veut dire qu'il était utilisé pour annoter des ressources différentes de domaine et de signification. Autrement dit, être présent dans plusieurs clusters signifie que le tag est ambigu, et il représente autant de sens que le nombre des clusters auxquels il appartient. Inspirés par les travaux antérieurs et après avoir eu des difficultés par rapport à la génération des clusters, nous avons élaboré l'algorithme 2, un nouvel algorithme de clustering flou désigné aux folksonomies. Nous décrivons tout d'abord l'algorithme puis, nous discutons de ses points forts et comment il répond aux problèmes d'autres algorithmes.

L'algorithme commence par ajouter le premier tag dans la liste de généralité générée précédemment comme le premier centre (ligne 1), ensuite il initialise le nombre de clusters "c" (ligne 2). Puis, il choisit d'autres centres à partir de la liste de généralité des tags. Ce choix est basé sur un seuil de similarité pour assurer de maximiser les distances entre les centres (lignes 4-10). Une fois tous les centres choisis, l'algorithme calcule la matrice d'appartenance des tags aux clusters qu'il rend en sortie avec le nombre de clusters générés.

Ce nouvel algorithme permet de surmonter les limitations de Fuzzy C-means (FCM) [Bezdek, 1981] et de plusieurs algorithmes de clustering. Nous citons ces limitations ci-dessous.

Algorithm 2 Identification de contexte et désambiguïsation des tags

Require: **NCDU[n][n]** {la matrice de similarité des tags}
Require: $L_{generality}$ {list des tags t_1,...,t_n dans un ordre décroissant de leur degré de généralité mesuré par FDU}
Require: **min_sim** {seuil pour choisir les centres}

```
centers [0] := L_generality[0] {le premier tag de la liste de généralité est le premier centre}
c=1 ;{Le nombre actuel des clusters}
i=1 ; {indice de la liste de généralité}
while i<L_generality.size do
  if NCDU[L_generality[i]][allcenters]<min_sim then
    c+ =1 ;
    centers[c] :=L_generality[i] ;
  end if
  i+=1 ;{un autre tag de L_generality}
end while
for i = 0 to t do
  for k = 0 to c do
    U=NCDU[i[k] ; {calcul des valeurs d'appartenance}
  end for
end for
return (c, U)
```

1. *Le nombre de clusters « c » doit être prédéfini :* le principal inconvénient du FCM et de la plupart des algorithmes de clustering est l'obligation de définir un nombre fixe de clusters. Cette tâche est très difficile, quelle que soit la nature des données, et plus difficile dans notre cas, car définir combien de clusters peuvent être générés lors du clustering d'un ensemble de tags dans une folksonomie n'est pas une tâche triviale.

2. La valeur du degré de flou « m » dans FCM est parmi les entrée de l'algorithme ; dans le cas de désambigüisation des tags, ce paramètre représente combien un tag ambigu peut avoir de significations différentes. La valeur de ce paramètre est souvent égale à 2 pour l'algorithme FCM, ce qui signifie dans notre cas, qu'aucun des tags ne peut avoir plus de deux sens différents, et tous les tags ambigus ont 2 significations différentes.

3. Dans tous les algorithmes de clustering l'attribution des tags aux clusters est basée sur la distance entre ces tags et les centres des clusters qui se calculent à partir des formules différentes, mais ces centres ne sont pas des tags et n'ont aucune signification pour la folksonomie, donc le contexte

d'un tag basé sur ses voisins dans le même cluster, y compris les tags bruyants. Ces derniers influence le résultat d'identification de contexte des tags.

4. Les résultats de clustering sont très sensibles aux initialisations, comme les centres et la matrice d'appartenance et un bon choix pour ces suppositions n'est pas évident.
5. Les algorithmes de clustering sont bien souvent très gourmands en ressources de calcul et prennent dès lors beaucoup de temps à s'exécuter.

Notre nouveau algorithme permet de surmonter toutes ces lacunes qu'il décide par lui-même de définir le nombre de clusters. Il s'agit d'ajouter de nouveaux centres tant qu'il existe des tags non classifiés, tout en assurant de garder une distance minimale entre les clusters. Aussi, l'algorithme ne nécessite aucune valeur du paramètre de degré de flou, donc il extrait le nombre effectif des sens disposant un tag ambiguï selon les valeurs de la similarité entre le tag et les centres des clusters associés. En outre, le contexte extrait par notre algorithme pour le tag est clairement défini puisque les centres des clusters sont un ensemble de tags appartenant à la folksonomie, de sorte qu'ils peuvent être considérés comme les définitions ou des significations de clusters. De plus, contrairement à d'autres algorithmes de clustering, ce nouvel algorithme ne nécessite aucune initialisation des entrées donc les résultats restent inchangeables avec la réexécution de l'algorithme tant que nous utilisons le même dataset. Enfin, ce nouvel algorithme est simple et performant par rapport aux autres algorithmes.

À la fin de l'étape d'identification de contexte et la désambigüisation, nous avons un ensemble de clusters flous. Les tags qui appartiennent à l'intersection sont ambigus, et le nombre de leurs significations est le nombre de clusters associés. Une fois cette étape réalisée, nous pouvons générer la hiérarchie des tags. Cette tâche est discutée dans la section suivante.

4.5 Identification de la sémantique des tags

Dans cette étape, nous induisons une organisation hiérarchique de l'espace des tags initialement plat. cette hiérarchie capture la sémantique et la diversité

des connaissances partagées. La majorité des approches associant des entités sémantiques aux tags s'appuient sur des techniques d'appariement de chaînes de caractères. Ces techniques visent à trouver des concepts correspondant dans l'ontologie candidate et ensuite utiliser le contexte du tag pour choisir le concept qui décrit le mieux son sens. Toutefois, cette activité implique le passage d'un espace plat, c'est-à-dire sans hiérarchies, sur le côté, folksonomie, à un espace hiérarchique dans le côté de l'ontologie. Quelques recherches comme celui de Angeletou et al. [Angeletou, 2008] abordent ce problème en associant les tags au départ aux WorldNet Synsets, et puis la structure hiérarchique des Synsets est comparée aux ontologies.

Nous suggérons, comme une alternative, de créer des hiérarchies de tags basées sur les informations de la folksonomie uniquement. Notre approche améliore l'approche de Heymann et al [Heymann, 2006]. Dans une étude comparative réalisée par [Strohmaier, 2012], il a été prouvé que l'algorithme de Heymann surpasse tous les algorithmes introduits dans l'étude. C'est pourquoi nous l'avons choisi comme le fondement de notre contribution. Nous utilisons des principes similaires à l'algorithme de Heymann. Cependant, notre nouvel algorithme (algorithme 3) offre les avantages suivants :

1. Au lieu d'utiliser la mesure de généralité par le degré de centralité du tag dans le réseau de cooccurrence tag-tag [Hoser, 2012] comme dans [Benz, 2010], nous utilisons FDU comme mesure de généralité. Ainsi, nous prenons en compte la dimension de l'utilisateur.
2. Les tags sous la racine sont les centres générés à l'étape d'identification de contexte et de désambigüisation.

L'algorithme commence par un arbre avec un seul nœud « racine » qui représente le sommet de l'arbre (ligne1). les centres générés par l'algorithme de clustering sont ajoutés à l'arborscence directement sous la racine, après avoir été supprimés de la liste de généralité (lignes 2 à 5)(pour éviter de les traiter une deuxième fois). Chaque tag est ajouté ensuite dans l'ordre décroissant de sa généralité. L'algorithme ajoute chaque tag au nœud le plus similaire dans l'arbre, si leur degré de similarité est supérieur à un seuil de similarité (lignes 7-18), sinon il est ajouté sous la racine (lignes 19 et 20). Dans le cas des tags ambigus, l'algorithme répète étapes 10-23 pour chacun de ses sens.

Algorithm 3 Identification de la sémantique des tags

Require: **nc**[t_1],...,**nc**[t_n] {le nombre des clusters où appartiennent les tags t_1,...,t_n }
Require: $L_{generality}$ {liste des tags t_1,...,t_n dans un ordre décroissant de leur degré de généralité mesuré par FDU}
Require: **min_sim** {seuil pour choisir ajouter les tags à la hiérarchie}
Require: **c** {le nombre des clusters générés}
Require: **centers**[**c**] {centres générés à l'étape de clustering}

```
1: Hierarchy← <root> ; {la racine de la hiérarchie est appelée "root" }
2: for i = 1 to c do
3:    Hierarchy← centers[i] ;
4:    L_generality.Remove(centers[i]) ; {supprimer le tag de la liste de généralité }
5: end for
6: for i = 0 to |(L_generality| do
7:    t_i ←L_generality[i] ;
8:    k=nc[t_i] ;
9:    maxCandidateVal = 0 ;
10:   repeat
11:      for all t_j ∈ getVertices(Hierarchy) do
12:         if NCDU(t_i,t_j)> maxCandidateVal then
13:            maxCandidateVal = NCDU(t_i,t_j)) ;
14:            maxCandidate =t_j ;
15:         end if
16:      end for
17:      if maxCandidateVal>min_sim then
18:      Hierarchy ← Hierarchy ⋃ <maxCandidate,(t_i >;
19:      else
20:         Hierarchy ← Hierarchy ⋃ <root,t_i> ; {t_i est ajouté à la hiérarchie sous
            la racine}
21:      end if
22:      k=k-1 ;
23:   until k=0 {tous les différents sens de t_i sont traités }
24: end for
25: Return (Hierarchy)
```

4.6 Conclusion

Dans ce chapitre, nous avons introduit notre approche pour l'extraction des structures ontologiques à partir des folksonomies, l'approche comporte plusieurs étapes en commençant par le prétraitement des tags pour les nettoyer et avoir des résultats plus précis, ensuite les tags nettoyés sont préparés pour effectuer les

traitements nécessaires pour arriver au résultat final sous forme d'une hiérarchie de tags. Cette préparation consiste à créer une liste des tags triés par un ordre décroissant de leur degré de généralité, à ce niveau-là nous avons contribué par une nouvelle mesure de généralité que nous appelons FDU. En outre une matrice carrée tag- tag représentant les similarités entre les tags est générée en se basant d'abord sur une nouvelle représentation de l'espace des tags et surtout sur une nouvelle mesure de généralité que nous appelons NCDU. Par la suite, cette liste, matrice, et le dataset nettoyé sont utilisés pour identifier le contexte des tags et pour pouvoir détecter et désambigüiser les tags ambigus. Cela s'effectue d'une manière aussi efficace que performante et ce, grâce au nouvel algorithme de clustuering que nous avons proposé. D'ailleurs l'idée d'utiliser un clustering flou a beaucoup réduit et simplifier la tâche, de plus l'indépendance de toute initialisation qui caractérise notre algorithme a beaucoup amélioré les résultats. Dans la dernière phase de notre approche nous avons amélioré et adapté un algorithme déjà existant qui génère une hiérarchie à partir d'un dataset [Heymann, 2006], les modifications que nous avons apportées ont aussi amélioré les résultats que nous discutons dans le chapitre suivant.

Chapitre 5

Expérimentation et évaluation

5.1 Introduction

Alors que nous avons détaillé dans le précédent chapitre notre approche pour l'extraction d'ontologies à partir des folksonomies, nous réalisons par la suite, une série d'expériences afin d'évaluer notre approche. Tout d'abord nous quantifions l'influence des mesures proposées (degré de similarité et de généralité des tags) sur la sémantique émergente de la folksonomie, ensuite nous évaluons la capacité du nouvel algorithme de clustering pour définir le contexte des tags et détecter ceux qui sont ambigus.

Avant d'expliquer chaque étape de l'expérimentation, nous fournissons dans la section suivante des détails sur les données utilisés.

5.2 DataSet

Nous avons utilisé pour les besoins de notre recherche des données de Flickr [28]. Le dataset a été collecté durant la période du 24 novembre au 31 décembre 2005. Il contenait à l'origine $| U |= 111.\ 920$ utilisateurs, $| R | = 3.253.390$ ressources, et $| T |= 374.076$ tags. Dans l'étape de nettoyage de notre approche, nous choisissons les tags les plus pertinents en se basant sur leur fréquence et leur signification. Il en est résulté un ensemble de données d'environ 18 329 images. Le vocabulaire associé a plus 6 798 termes choisis par 1, 904 utilisateurs.

Nous utilisons également un dataset du système social bookmarking Del.icio.us [29], recueilli en novembre 2006 le total des données comprend 663, 950 utilisateurs, 2, 398, 483 tags et 18, 775, 470 ressources. Après le nettoyage des données nous avons abouti à 23, 643 utilisateurs, 48, 135 ressources et 8, 568 tags.

5.3 Préparation des tags

Dans cette étape, nous générons une liste de tags classés selon leur degré de généralité. La mesure de généralité mise en œuvre est celle que nous avons

28. http ://west.uni-koblenz.de/Research/DataSets/PINTSExperimentsDataSets
29. http ://west.uni-koblenz.de/Research/DataSets/PINTSExperimentsDataSets

proposée (FDU). les Tableaux 5.1 et 5.2 décrivent des extraits de cette liste pour les datasets Flickr et Del.icio.us respectivement.

Tag	Food	Music	Sport	Restaurant	Japan	Transport	Friend
Generality degree	4561	1114	856	373	366	276	265

TABLE 5.1 – Extrait de la liste de généralité générée pour Flickr, mesurée par FDU

Tag	Web	Design	Html	Blog	Search	News	Art	Color
Generality degree	8680	4720	2950	2373	2280	2200	1290	1130

TABLE 5.2 – Extrait de la liste de généralité générée pour Delicous, mesurée par FDU

Dans cette étape, nous générons également une matrice de similarités entre les tags. Cette matrice est basée sur la mesure NCDU. Cette mesure de similarité donne des résultats plus précis que d'autres mesures telles que la coocurrence, CDU [Marouf, 2013] et Cosine. À titre d'exemple, nous donnons dans le tableau 5.3, un extrait de la matrice NCDU et dans les tableaux 5.4,5.5, 5.6, 5 termes les plus similaires à certains tags selon les mesures NCDU, coocurrece, et Cosine.

	Food	Girl	Airline	Music
Actress	0	0,66666667	0	0,33333333
Adventure	0,2	0	0	0
Advertise	0,58333333	0	0	0,08333333
Airplane	0,6097561	0	0,5	0,04878049
Banjo	0	0	0	0,71428571
Baseball	0,09638554	0,02409639	0	0
Guitar	0	0,02020202	0	0,83458647

TABLE 5.3 – Extrait de la matrice NCDU

5.4 Identification de contexte et désambiguïsation

Comme fut expliqué dans le chapitre précédent, les tâches d'identification de contexte et de la désambiguïsation sont effectuées en une seule étape en appliquant

	(1)	(2)	(3)	(4)	(5)
Architecture	Rio	Skyscraper	Historic	Building	Historical
Canada	Quebec	Edmonton	Ontario	Toronto	Alberta
Art	Actress	Rio	Craft	Fine	Write
Fun	Food	Lifestyle	Teenager	Zoo	Sad
Cat	Kitten	Food	Dolphin	Amazon	Pet
Band	Music	Ballroom	Gig	Guitarist	Nightclub
Famous	Actress	Food	Interestingness	Celebrity	Sport
Car	Nissan	Automotive	Ford	Mustang	Auto

TABLE 5.4 – 5 termes les plus similaires à certains tags celon NCDU

	(1)	(2)	(3)	(4)	(5)
Architecture	Animal	Flower	Cat	Tree	Dog
Canada	Food	Animal	Flower	Cat	Mountain
Art	Animal	Music	Flower	Tree	Cat
Fun	Music	Animal	Friend	Cat	Flower
Cat	Animal	Flower	Tree	Dog	Food
Band	Music	Live	Rock	Nightclub	Concert
Famous	Food	Yahoo	Interestingness	Brasil	Sport
Car	Day	Race	Bahrain	Animal	Prom

TABLE 5.5 – 5 termes les plus similaires à certains tags selon la mesure de co-ocurrence

	(1)	(2)	(3)	(4)	(5)
Architecture	Red	Fish	Travel	January	California
Canada	Photo	Fish	Doughnut	Drink	Travel
Art	Tokyo	June	Topv	Reflection	Window
Fun	People	July	Team	Beer	Thai
Cat	House	France	Grill	Cookie	July
Band	Travel	China	Macro	Birthday	Red
Famous	Scott	South	Brighton	Dance	Paprika
Car	April	Travel	Red	Birthday	California

TABLE 5.6 – 5 termes les plus similaires à certains tags selon la mesure de Cosine

un nouvel algorithme de clustering. Les figures 5.1, 5.2 et 5.3 représentent des exemples de tags ambiguës « player » et « adventure », et « design » tels qu'ils sont détectés par l'algorithme.

Le tableau 5.7 présente d'autres exemples de tags ambigus détectés par le nouvel algorithme.

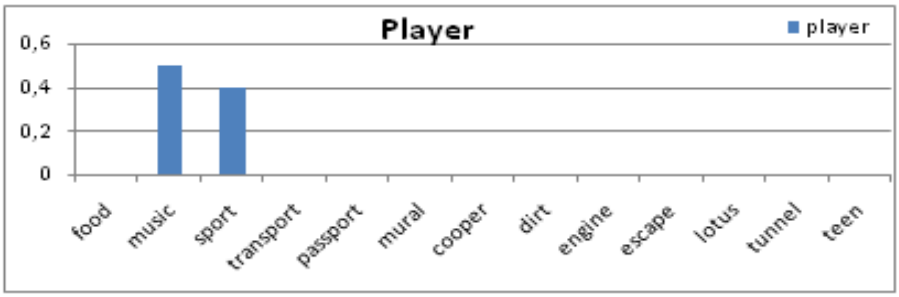

FIGURE 5.1 – Le tag ambigu player.

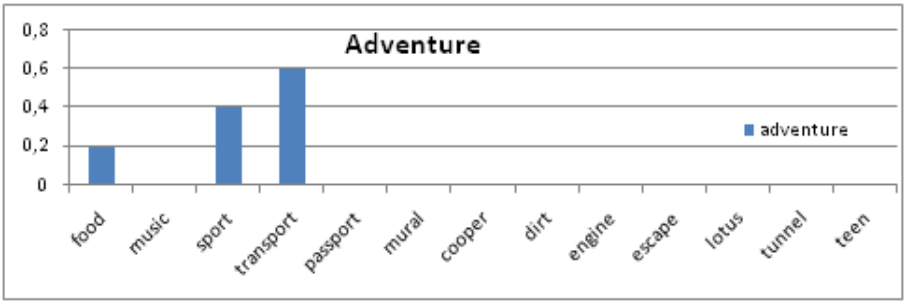

FIGURE 5.2 – Le tag ambigu adventure.

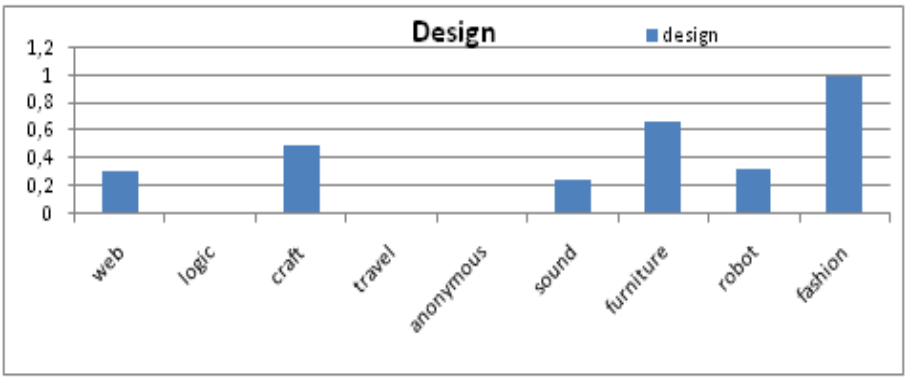

FIGURE 5.3 – Le tag ambigu design.

Tag	centres des cluster auquels appartient le tag
Track	Sport / Transport
Entertainment	Music/ Sport
Swing	Music /Sport
Speed	Sport / Transport / Engine
Music	Mp / Loop/ Sound
Spider	Web / Robot
Video	Multiplication / Sound

TABLE 5.7 – 5 termes les plus similaires à certains tags selon la mesure de Cosine

5.5 Identification de la sémantique des tags

Dans cette étape, nous avons généré des hiérarchies des tags à l'aide de l'algorithme décrit au chapite précédent. Les figures 5.4 et 5.5 illustrent des extraits de ces hiérarchies des tags de Flickr et Delicious respectivement.

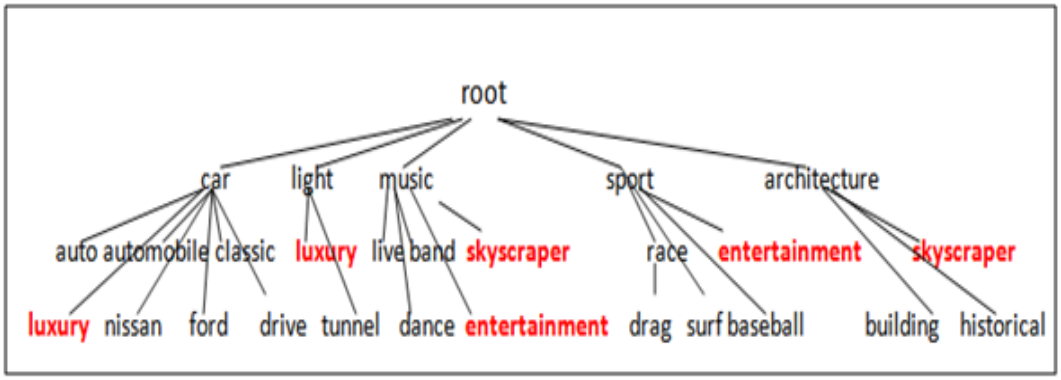

FIGURE 5.4 – Extrait de la hiérarchie des tags de Flickr (les noeuds rouges sont des tags ambigus).

5.6 Evaluation

Afin d'évaluer la qualité de la hiérarchie générée, nous la comparons avec les schémas de catégorisation construits manuellement à partir WordNet et Wikipedia. Malgré cela, il n'est pas évident de trouver un score de similarité valide pour deux structures hiérarchiques. Pour ce faire, nous utilisons la mesures TO (Taxonomic Overlap) (proposées par [Maedche, 2002a], et la mesure F-measure , initialement introduit dans la recherche d'information. Le principe est de trouver

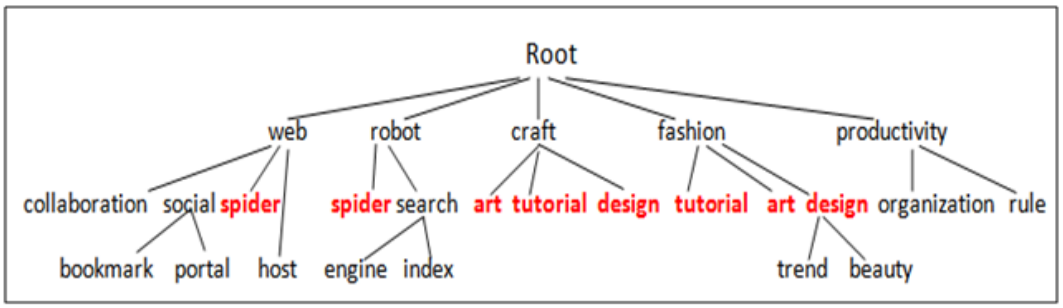

FIGURE 5.5 – Extrait de la hiérarchie des tags de delicious (les noeuds rouges sont des tags ambigus)..

un concept présent dans les deux hiérarchies et extraire deux extraits des deux ontologies contenant ce concept, alors la similarité entre les deux hiérarchies est dependante de la similitude des deux extraits.

Basés sur ces mesures et les deux ontologies, nous avons effectué plusieurs expériences pour évaluer la qualité de notre approche, À cette fin, nous avons généré des taxonomies basées sur des mesures de similarités telles que la cooccurrence, cosine et NCDU. Les résultats de la comparaison sont représentés sur la figure 5.6 pour la mesure TO, et sur la figure 5.7 pour F-mesure.

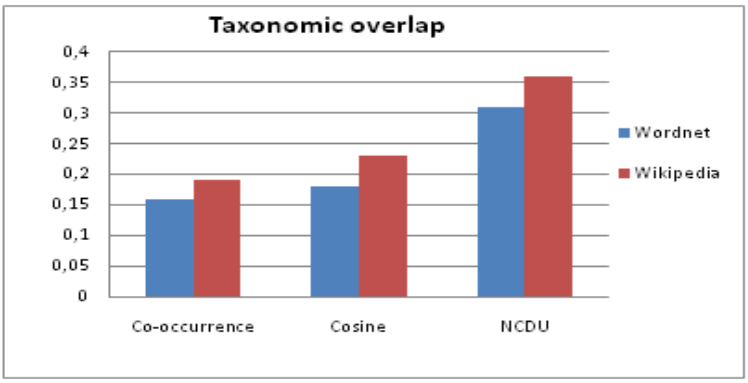

FIGURE 5.6 – Comparaison basée sur la mesure TO entre les structures hiérarchiques générées et les ontologies de référence WordNet et Wikipedia.

Nous avons également testé la qualité de la mesure de similarité proposée (NCDU). Pour cette raison, nous avons utilisé le tau de Kendall qui est une technique statistique utilisée pour mesurer la similarité entre deux ensembles de données. Dans nos expérimentations, nous calculons les corrélations entre les

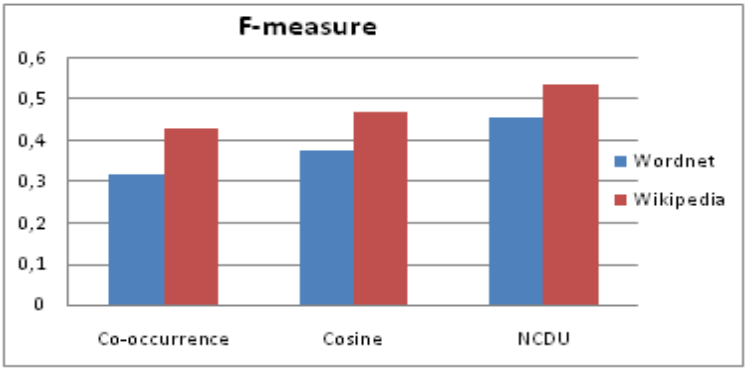

FIGURE 5.7 – Comparaison basée sur F-measure entre les structures hiérarchiques générées et les ontologies de référence WordNet et Wikipedia.

valeurs de similarité basées sur les mesure de cooccurrence, cosine et NCDU d'une part et les valeurs de similarité fournies par WordNet d'autre part. La formule de τ est comme suit :

$$\tau = \frac{n}{\frac{1}{2}p(p-1)} \tag{5.1}$$

Où n = (nombre total de paires concordant)-(nombre total de paires discordant), et p est le nombre total de paires. Les paires concordantes sont définies comme suit : Étant donné deux paires $p_1 = (t_1;t_2)$ et $p_2 = (t_3;t_4)$. P_1 et p_2 sont concordantes si $t_1 > t_3$ et $t_2 > t_4$, ou si $t_1 < t_3$ et $t_2 < t_4$. les valeurs de τ varient entre -1 et 1, où -1 signifie que toutes les paires sont discordantes et 1 signifie que tous les paires sont concordantes. la figure 5.8 montre le résultat de comparaison de notre nouvelle mesure de similarité avec les mesures coocurrence et cosine en utilisant la corrélation de Kendall entre leurs valeurs de similarité et les similarités générées par la référence WordNet.

Nous avons utilisé la même formule pour comparer la mesure de généralité proposée FDU avec la mesure de généralité basée sur la fréquence. Cependant, nous avons là une autre définition de paires concordantes et discordantes. Étant donné une paire de tags p = $(t_1 ; t_2)$, p est une paire concordante si $t_1 > t_2$ avec la mesure de généralité aussi bien que dans WordNet. La figure 5.9 illustre la comparaison entre ces mesures.

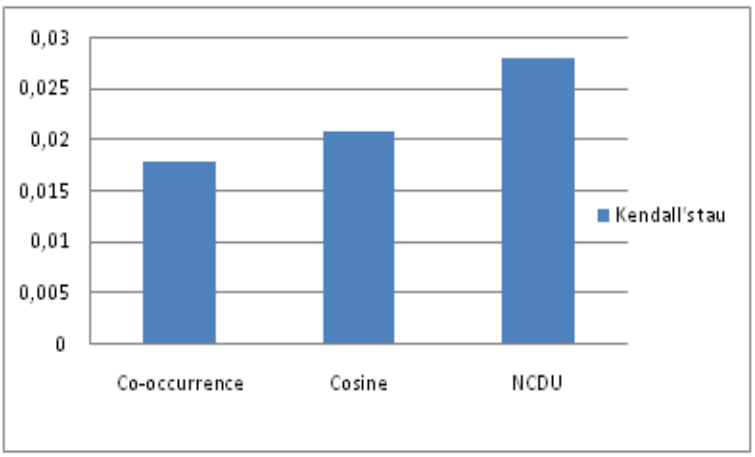

FIGURE 5.8 – Comparaison basée sur le tau de kendall entre les mesures de similarité.

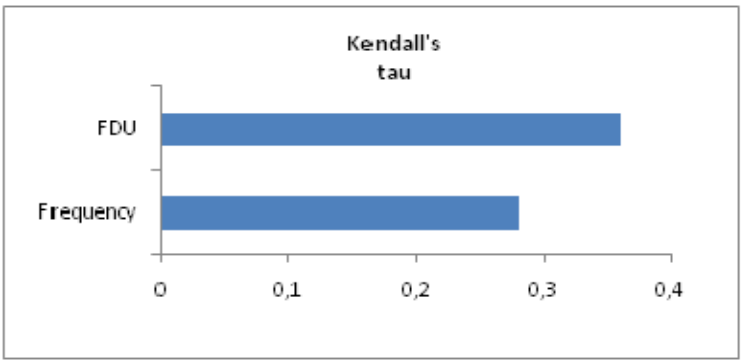

FIGURE 5.9 – Comparaison basée sur le tau de kendall entre les mesures de généralités.

L'objectif de l'évaluation d'un algorithme de clustering f est de rechercher des schémas d'agrégation, où minimiser l'inertie intra-cluster et maximiser l'inertie inter-cluster. À cette fin, nous avons utilisé le coefficient de partition (Partition Coefficient PC) [Bezdek, 1981] pour comparer notre algorithme avec l'algorithme FCM qui est l'algorithme le plus répandu parmi les algorithmes de clustering flous. La formule de calcule de PC est donnée par l'equations suivante.

$$PC = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^2 \tag{5.2}$$

Où n et k sont le nombre des tags et de clusters respectivement.

Nous effectuons une série de résultats expérimentaux pour comparer les performances de notre nouvel algorithme avec FCM. Les expériences sont effectuées sur un ordinateur personnel Intel core i3 2,4 GHz à 64 bits, 4Go de mémoire et 464 GO de disque dur. Il est évident que notre nouvel algorithme surpasse d'une façon remarquable l'algorithme FCM en terme de qualité des clusters générés et en terme de performances comme indiqué dans figure 5.10 et tableau 5.8.

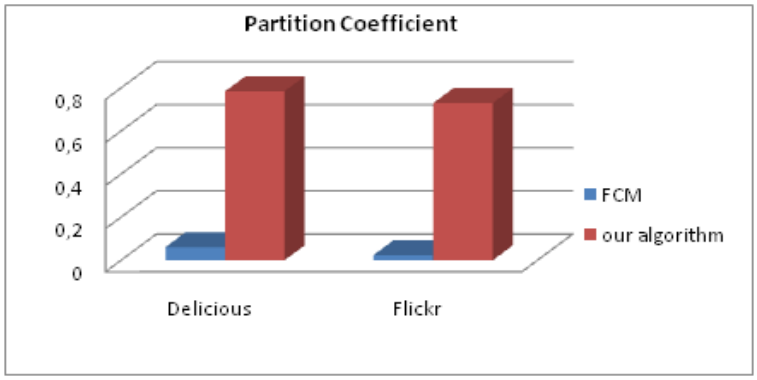

FIGURE 5.10 – Résultats de comparaison de la qualité de notre nouvel algorithme et celle de FCM.

	Delicious	Flickr
temps d'exécution de FCM en secondes	622.349	593.288
temps d'exécution de notre algorithme en secondes	1.214	1.035

TABLE 5.8 – Résultats de comparaison de la performance de notre nouvel algorithme et celle de FCM

5.7 Conclusion

Dans ce chapitre, nous avons présenté les résultats expérimentaux et quelques évaluations de notre approche. Le problème abordé dans la présente évaluation est à quel point notre approche en général et spécifiquement nos mesures de similarité et de généralité sont aptes à donner les mêmes résultats que les ontologies construites manuellement. Nos propositions génèrent des ontologies plus proches de celles utilisées dans l'étude (c.-à-d., Wordnet et Wikipedia). Cela revient au fait que notre mesure de similarité est plus mature et expressive que les autres mesures. En effet, elle implique tous les acteurs existants dans les systèmes de

collaboration. Un autre effet du succès de notre proposition est l'étape d'identification de contexte et de désambiguïsation basée sur un algorithme de clustering simple, efficace et indépendant de toute initialisation qui pourra affecter les résultats,

Chapitre 6

Conclusion et perspectives

6.1 Conclusion

En introduction de cette thèse, nous avons résumé la problématique scientifique motivant nos travaux de la manière suivante : Comment combiner les ontologies et les folksonomies afin de tirer profit de la richesse, la simplicité et le coût réduit des folksonomies issues d'outils du Web 2.0 pour la représentation et l'exploitation de connaissances formalisées selon les principes des ontologies. Ainsi, nous avons montré de quelle manière nous envisagions cette complémentarité qui permet de surmenter les difficultés rencontrés lors du développement des ontologies à savoir la lourdeur de la tâche, le besoin de l'expertise, le coût élevé, le besoin du consensus..., et en outre remédier aux inconvénients des folksonomies qui souffrent de l'emploi de vocabulaire non controlé ouvrant la possibilité de contenir des termes mal sélectionnés soit pour des raisons reliées à leurs syntaxe ou leur sémantique. Dans ce travail nous avons proposé une démarche pour avoir cette complémentarité, notre façon de combiner les ontologies et les folksonomies consiste à extraire des structures ontologiques à partir des folksonomies. Dans cette thèse ces structures ontologiques sont réduites aux relations taxonomiques qui existent entre les tags des folksonomies. Ainsi les principales contributions de cette thèse sont :

1. Une nouvelle representation de la folksonomy basée sur une matrice binaire tag-tag au lieu de resource- resource dans le but d'améliorer la performance des experimentations
2. Une nouvelle mesure de similarité plus simple et plus précise pour extraire les relations entre les tags
3. Une nouvelle mesure de généralité simple et efficace pour calculer le degré d'abstraction des tags, cette mesure est utilisée pour extraire la hierarchie des tags dans la folksonomie.
4. La prise en compte et la résolution du problème d'ambiguïté lors du processus de generation de l'ontologie légère à partir de la folksonomie. Cela a permis de retourner des resultats plus précis par rapport aux mecanismes de recherche traditionnels.
5. Nous avons tiré profit de la propriété du clustering flou qui permet au cluster de se chevaucher pour distinguer les tags ambigus et pour definir leurs différents sens à la fois.

6. Nous avons proposé un nouvel algorithme de clustering flou plus simple, plus efficace, et ne nécessitant aucune donnée préalablement définie comme le nombre des clusters à générer ou autre.

Notre recherche a passé par plusieurs étapes, la première idée que nous avons exploitée est de combiner des approches déjà existantes, et qui ont donné des résultats meilleurs par rapport à d'autres approches dans des évaluations faites par plusieurs chercheurs. Cela dans le but d'avoir un ensemble d'étapes pour l'extraction de la sémantique à partir des folksonomies. De se fait, la première version de notre approche présentait les étapes suivantes :

1. Nettoyage effectuée par un algorithme regroupant plusieurs tâches proposées dans des différents travaux de l'art.
2. Ensuite l'étape de préparation des tags qui avait des contributions aux niveaux de la mesure de similarité et de généralité mais qui utilise une technique d'aggrégation déjà exitante (la macro agrégation [Markines, 2009])
3. L'étape suivante c'est l'identification de contexte et la désambigüisation, nous avions une idée originale dans ce contexte qui consiste à utiliser un clustering flou pour effectuer les deux étapes en même temps ; le clustering flou permet de regrouper les éléments similaires dans des clusters et permet à ces éléments d'appartenir à plusieurs clusteurs. Nous avons exploiter cette idée pour cette étape en effectuant un clustering flou des tags de la folksonmie pour definir le contexte du tag t comme étant l'ensemble des tags du même cluster que t. avec le même algorithme nous avons pu détecté les tags ambigus et les désambigüiser ; un tag est reconnu ambigu s'il appartient à l'intersection de plusieurs clusters, et ses différents sens correpondent à ses différents contextes. Nous avions déjà cette idée mais nous avons utilisé un algortithme de clustering flou déjà existant
4. La dernière étape est l'identification de la sémantique des tag : pour cela nous avons utilisé un algorithme déjà existant pour la génération de la hierarchie [Heymann, 2006].

Cette première version de l'approche a été communiquée dans la conférence internationale AICCSA 2013 [Marouf, 2013].

Ensuite nous avons apporté quelque amélioration sur l'approche pour arriver à la version finale, en proposant :

1. Une nouvelle agrégation
2. Une deuxième mesure de similarité
3. Un nouvel algorithme de clustering flou
4. Une amélioration de l'algorithme de Heymann et al [Heymann, 2006] pour la génération de la hiérarchie à partir de la folksonomie.

L'ontologie générée peut être employée pour améliorer plusieurs tâches telles que l'enrichissement et l'évolution d'ontologie par ajout de nouveaux concepts créés par la communauté. Elle peut également être utilisée dans la recommandation des tags pour guider le processus du tagging en proposant des tags plus contrôlés. En outre, cette ontologie peut être utilisée dans l'expansion des requêtes afin d'améliorer les performances de la recherche d'information.

Cette dernière version était publiée dans le journal IJITCS (International Journal of Information Technology and Computer Science)[Marouf, 2014a], et dans le journal IJIIT (International Journal of Intelligent Information Technologies) [Marouf, 2014b].

D'autres perspectives et améliorations pourraient être apportées à notre approche et que nous survolerons dans la section suivante

6.2 Perspectives

À l'issue de cette thèse, différentes perspectives de recherche venant dans la continuité des travaux présentés dans cette thèse s'offrent à nous. Nous souhaitons ainsi axer une partie de nos travaux futurs autour des problématiques suivantes :

1. Amélioration de l'approche en découvrant des relations non taxonomiques par la detection des tag représentant des verbes et leurs tags liés.
2. Évaluation extrinsèque la qualité de nos résultats en les intégrant dans le cadre de diverses tâches telles que la désambiguisation des termes, la visualisation des résultats et de l'évolution de l'ontologie.

3. Amélioration de l'appliction developpée pour qu'elle soit un outil d'extraction des ontologies à partir des folksonomies.
4. Validation de l'ontologie extraite

Bibliographie

[Abbassi, 2007] Abbassi R., Staab S. & Cimiano P. (2007) " Organizing Resources on Tagging Systems using T-ORG " . In 4th European Semantic Web Conference. pp. 97-110.

[Anderson, 2006] Anderson C. (2006). " The Long Tail : Why the Future of Business is Selling Less of More ". Hyperion. 256p.

[Angeletou, 2008] Angeletou S., Sabou M. & Motta E. (2008). " Semantically Enriching Folksonomies with FLOR ". In 1st International Workshop on Collective Semantics : Collective Intelligence & the Semantic Web (CISWeb 2008). Tenerife, Spain.

[Angeletou, 2009] Angeletou S., Sabou M., & Motta E. (2009). " Improving Folksonomies Using Formal Knowledge : A Case Study on Search" . In 4th Asian Semantic Web Conference, pp. 276-290.

[Au Yeung, 2009] Au Yeung C. M., Gibbins N., & Shadbolt N. (2009). "Contextualizing Tags in Collaborative Tagging Systems ". In : 20th Conference on Hypertext and Hypermedia, pp. 251-260.

[Auer 2007] Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R. & Ives Z. (2007). " DBpedia : A Nucleus for a Web of Open Data ". In 6th International Semantic Web Conference. pp. 722-735. Busan, Korea.

[Benz, 2010] Benz D., Hotho A. & Stumme G. (2010). " Semantics Made by You and Me : Self-emerging Ontologies can Capture the Diversity of Shared Knowledge ". In : Proceedings of the 2nd Web Science Conference (WebSci10). Raleigh, NC, USA.

[Benz, 2011] Benz B., Körner C., Hotho A., Stumme G., & Strohmaier M. (2011). " One tag to bind them all : Measuring term abstractness in social metadata ". In : Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011). pp. 360-374. Heraklion, Crete, Greece.

[Berendt, 2007] Berendt B & Christoph Hanser. (2007). " Tags are not metadata, but "just more content" - to some people ". In Proceedings of the First International Conference on Weblogs and Social Media (ICWSM2007). Colorado

[Berners-Lee, 1990] Berners-Lee T. & Cailliau, R. (1990). "WorldWideWeb : Proposal for a HyperText Project". http ://www.w3.org/Proposal.html

[Berners-Lee, 2000] [Berners-Lee, 2000] Berners-Lee T. (2000). "Weaving the Web". Orion Business Books.

[Bezdek, 1981] Bezdek J.C. (1981). "Pattern Recognition with Fuzzy Objective Function Algorithms" . Plenum Press. New York.

[Bo Leuf, 2001] Bo L. & Ward C. (2001). "The Wiki Way : Collaboration and Sharing on the Internet" . Addison-Wesley Professional.

[Borst, 1997] Borst p., Akkermans H. & Top J. (1997). " Engineering ontologies". International Journal of Human-Computer Studies 46 (2-3), pp. 365-406

[Boullier, 2008] Boullier D. (2008). "Politiques plurielles des architectures d'Internet" . Cahier Sens Public, L'internet entre savoirs, espaces publics et monopoles. N 7-8. pp. 177-202.

[Brandes, 2007] Brandes U. & Pich C. (2007). "Centrality estimation in large networks". International journal Bifurcation and Chaos , 17(7), pp. 2303-2318.

[Cantador 2008] Cantador I., Szomszor M., Alani H., Fernandez M. & Castells P. (2008). "Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations" . In 1st International Workshop on Collective Semantics, (CISWeb 2008), Tenerife, Spain.

[Cattuto, 2008] Cattuto C., Benz D., Hotho A. & Stumme G. (2008). "Semantic grounding of tag relatedness in social bookmarking systems". In The Semantic Web - ISWC 2008 , edited by Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin and Krishnaprasad Thirunarayan , pp. 615–631.

[Cimiano, 2006] Cimiano P. (2006). "Ontology Learning and Population from Text : Algorithms, Evaluation and Applications". Secaucus, NJ, USA : Springer-Verlag New York.

[Crepel, 2011] Crepel M. (2011). "Tagging et folksonomies : pragmatique de l'orientation sur le Web". Computers and Society. Université de Rennes 2. French.

[Deuff, 2007] Le Deuff O. (2007). "Folksonomies : Les usagers indexent le web" , Bulletin des bibliotheques de France, 51(4).

[Ding, 2009] Ding Y., Jacob E.K., Fried M., et al. (2009). " Upper tag ontology for integrating social tagging data", Journal of the American Society for Information Science and Technology. 61(3), pp. 505521.

[Ertz, 2006] Ertzscheid O. & Gallezot G. (2006). "Etude exploratoire des pratiques d'indexation sociale comme une renégociation des espaces documentaires. Vers un nouveau big bang documentaire ? " in Chartron G. & Broudoux, E. (Eds). Document numérique et société, ADBS Éditions, 2006 en ligne sur http ://archivesic.ccsd.cnrs.fr/

[Firth, 1957] Firth J. (1957). "A synopsis of linguistic theory 1930-1955". Studies in linguistic analysis, pp. 132.

[Gandon, 2009] Gandon F., Limpens D., Monnin A. et al. (2009). "Nicetag ontology : tags as named graphs". SNI09 - Social Networks Interoperability, 1st International Workshop at the 4th Asian Semantic Web Conference (ASWC 2009), Shanghai, China, December 2009. Disppnible sur http ://spin.nicta.org.au/SNI2009/SNI2009/Program files/nicetag full.pdf

[Garcia-Silva 2009] Garcia-Silva A., Szomszor M., Alani H. & Corcho O. (2009). "Preliminary Results in Tag Disambiguation using DBpedia" . In 1st International Workshop in Collective Knowledge Capturing and Representation (CKCaR09), California, USA.

[Giannakidou 2008] Giannakidou E., Koutsonikola V., Vakali A. & Kompatsiaris Y. (2008). "Co-Clustering Tags and Social Data Sources". In 9th International Conference On Web-Age Information Management. Los Alamitos, USA. pp. 117-324.

[Golder, 2006] Golder S. A. & Huberman B. A. (2006). "Usage patterns of collaborative tagging systems" . Journal of Information Science.32 (2). pp. 198-208

[Gruber, 1993] Gruber T. R. (1993). "A translation approach to portable ontology specifications ". Journal of Knowledge Acquisition. 5(2). pp. 199-220.

[Guarino, 1993] Guarino N. (1997). "Semantic Matching : Formal Ontological Distinctions for Information Organization, Extraction, and Integration " . International Summer School on Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology. Springer-Verlag. pp. 139-170.

[Guy, 2006] Guy M. & Tonkin E. (2006). " Folksonomies Tidying up Tags ? ". D-Lib Magazine. 12(1). http ://www.dlib.org/dlib/january06/guy/01guy.html

[Hamasaki, 2007] Hamasaki M., Matsuo Y., Nisimura T. & Takeda, H. (2007). "Ontology Extraction using Social Network". In International Workshop on Semantic Web for Collaborative Knowledge Acquisition. India. pp.1-6.

[Heymann, 2006] Heymann P. & Garcia-Molina H. (2006). "Collaborative creation of communal hierarchical taxonomies in social tagging systems". Technical Report . Computer Science Department, Standford University.

[Heymann, 2008] Heymann P., Ramage D., Garcia-Molina H. (2008). "Social tag prediction". In 31st annual International ACM SIGIR conference on Research and development in information retrieval SIGIR '08. ACM, New York, NY, USA .pp. 531-538

[Hoser, 2012] Hoser B., Hotho A., Jaschke R., Schmitz C., & Stumme G. (2006). "Semantic network analysis of ontologies". In European Semantic Web Conference, Budva, Montenegro. Pp. 514-529.

[Jäschke 2008] Jäschke R., Hotho A., Schmitz C., Ganter B. & Stumme G. (2008). "Discovering shared conceptualizations in folksonomies" , Journal of Web Semantics : Science, Services and Agents on the World Wide Web. 6 (1). pp. 38-53.

[Kennedy, 2007] Kennedy L., Naaman M., Ahern S., Nair R. & Rattenbury T. (2007). "How Flickr Helps us Make Sense of the World : Context and Content in Community-Contributed Media Collections" . In Proceedings of 15th International Conference on Multimedia MULTIMEDIA '07. Augsburg, Germany. pp. 631-640.

[Kim, 2008] Kim H., Scerri S., Breslin J., et al. (2008). " The state of the art in Tag Ontologies : A semantic model for tagging and folksonomies" . International Conference on Dublin Core and Metadata Applications DC8. Berlin. pp. 128-137.

[Knerr, 2007] knerr T. (2007). "Tagging Ontology - Towards a Common Ontology for Folksonomies ". Disponible sur : https ://code.google.com/p/tagont/downloads/detail ?name=TagOntPaper.pdf

[Koerner, 2010] Koerner C., Benz D., Strohmaier M., Hotho A. & Stumme G. (2010). "Stop thinking, start tagging - tag semantics emerge from collabora-

tive verbosity" . In : Proceedings of the 19th International World Wide Web Conference (WWW 2010), Raleigh, NC, USA. pp. 521-530.

[Lando, 2006] Lando p.(2006). "Conception et développement d'applications informatiques utilisant des ontologies : application aux EIAH ". Actes des Premières rencontres jeunes chercheurs en EIAH (RJC-EIAH'06), Évry, France.

[Laniado, 2007] Laniado D., Eynard D. & Colombetti M. (2007). " Using WordNet to turn a folksonomy into a hierarchy of concepts". In Semantic Web Application and Perspectives Fourth Italian Semantic Web Workshop. Bari, Italy. pp.192-201.

[Lehmann 1995] Lehmann F. & Wille. F.R. (1995). " A triadic approach to formal concept analysis". In the Third International Conference on Conceptual Structures : Applications, Implementation and Theory. pp. 32-43.

[Limpens, 2010] Limpens F. (2010). " Multi-points of view enrichment of folksonomies", Thèse sous la direction de GANDON F., INRIA Sophia Antipolis, Edelweiss and Computer Science Dpt of niversité Nice Sophia Antipolis. Disponible sur : http ://tel.archives-ouvertes.fr/tel-00530714/fr/

[Lin 2009] Lin H., Davis J. & Zhou Y. (2009). "An integrated approach to extracting ontological structures from folksonomies". In the 6th European Semantic Web Conference on The Semantic Web : Research and Applications. pp. 654-668.

[Maedche, 2002a] Maedche A. & Staab S. (2002). "Measuring similarity between ontologies". In the 13th International Conference on Knowledge Engineering and Knowledge Management, Ontologies and the Semantic Web EKAW 02 . London, UK. pp. 251263.

[Marchetti, 2007] Marchetti A. & Rossela A. (2007). " SemKey : A Semantic Collaborative Tagging System". In WWW 2007 Workshop on Tagging and Metadata for Social Information Organization. Banff, Canada.

[Markines, 2009] Markines B., Cattuto C., Menczer F., Ben D., Hotho A.& Stumme, G. (2009). "Evaluating Similarity Measures for Emergent Semantics of Social Tagging". In : 18 th International Conference on World Wide Web. pp. 641-650.

[Marlow, 2006] Marlow C., Naaman M., Boyd D. & Davis M. (2006). "HT06, tagging paper, taxonomy, Flickr, academic article, to read". In the seventeenth conference on Hypertext and hypermedia HYPERTEXT 06. pp. 3140.

[Marouf, 2013] Marouf Z. & Benslimane S. M. (2013). "Fuzzy clustering-based approach to derive hierarchical structures from folksonomies". In International conference on computer systems and applications AICCSA. Fes, Morocco. pp :1-8.

[Marouf, 2014a] Marouf Z., Benslimane S.M. (2014). "An Integrated Approach to Drive Ontological Structure from Folksonomie". In International Journal of Information Technology and Computer Science(IJITCS) . 6(12). pp.35-45.

[Marouf, 2014b] Marouf Z., Benslimane S.M. (2014). "Towards Ontological Structures Extraction from Folksonomies : An Efficient Fuzzy Clustering Approach" . In International Journal of Intelligent Information Technologies (IJIIT). 10(4). pp. 40-50.

[Mika, 2007] Mika, P.(2007). "Ontologies are us : A unified model of social networks and semantics". In Journal of Web Semantics. 5(1). pp. 5-15.

[NRC, 1995] National Research Council (U.S.). (1995) Committee on the Future of the Global Positioning System ; National Academy of Public Administration (1995). "The global positioning system : a shared national asset : recommendations for technical improvements and enhancements". National Academies Press. ISBN 0-309-05283-1., Chapter 1, p.16

[Orlean, 1994] Orlean A. (1994). "Analyse économique des conventions", Presses universitaires de France PU 1994F. Paris. Pp. 219-247.

[Pantel, 2002] Pantel P. & Lin D. (2002). "Document clustering with committees". In 25th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR02. Tampere, Finland. pp. 199-206.

[Passant, 2007] Passant A. (2007). " Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs" . In the first International Conference on Weblogs and Social Media ICWSM 07. Colorado. USA.

[Passant,2008] Passant A. & Laublet P.(2008). " Meaning Of A Tag : A collaborative approach to bridge the Gap Between Tagging and Linked Data" . In 17 th international world wide web conference linked data on the web workshop LDOW2008. Beijing. China.

[Quintarelli, 2007] Quintarelli E., Resmini A. & Rosati L. (2007). "FaceTag : integrating bottom-up and top-down classification in a social tagging system". In the International IA Summit. Las Vegas.

[Ronzano, 2008] Ronzano F., Marchetti A. & Tesconi M. (2008). "Tagpedia : a semantic reference to describe and search for Web resources" . In of The workshop Social Web and Knowledge Management of the World Wide Web Conference SWKM2008. Beijing.

[Schmitz 2006] Schmitz C., Hotho A., Jaschke R. & Stumme G. (2006). "Mining association rules in folksonomies" . In the 10th IFCS Conference, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg. pp 261270.

[Shepitsen, 2008] Shepitsen A., Gemmell J., Mobasher B. & Burke R. (2008). "Personalized Recommendation in Social Tagging Systems using Hierarchical Clustering". In ACM Conference on Recommender Systems. pp 259-266.

[Shirky, 2007] Shirky C. (2005). "Ontology is Overrated : Categories, Links, and Tags". Clay Shirky's Writings About the Internet . Economics & Culture, Media & Community. Disponible sur http ://www.shirky.com/writings/ontology_overrated.html

[Specia, 2007] Specia L.& Motta E. (2007). "Integrating Folksonomies with the Semantic Web" . In 4th European Semantic Web Conference ESWC. Innsbruck, Austria. pp. 624-639

[Spiteri, 2008] Spiteri L. (2008). "Editorial : Folksonomies, the Web and Search Engines", in the nternational scholarly open access journal Webology. 5(3). editorial 17. Disponible sur http ://www.webology.org/2008/v5n3/editorial17.html

[Spivack, 2009] Spivack N. (2009). "Web Evolution". Presentation sur Slideshare. Disponible sur http ://www.slideshare.net/novaspivack/web-evolution-nova-spivack-twine

[Strohmaier, 2012] Strohmaier M. , Helic D., Benz D., Körner C. & Kern R. (2012). "Evaluation of folksonomy induction algorithms". In the journal of ACM Transactions on Intelligent Systems and Technology (TIST). 3(4). pp. 74 :1-74 :22.

[Taylor, 1999] Taylor A. G. (1999). "The Organization of Information" . In Libraries Unlimited. Englewood, CO, USA. 512 p.

[Tim O'Reilly, 2005] O'Reilly T. (2005). "O'Reilly Network : What Is Web 2.0 : Design Patterns and Business Models for the Next Generation of

Software" . In Communications & Strategies. No. 1. p. 17. Disponible sur http ://www.oreillynet. com/lpt/a/6228

[Wasserman, 1994] Wasserman S. & Faust K. (1994). "Social network analysis : Methods and applications" . In Cambridge University Press. 825 p.

[Weinberger, 2007] Weinberger D. (2007). "Everything is miscellaneous : The power of the new digital disorder", Henry Holt and Company, Information Research. 12(4), review no. R269. 278 p.

[Weinberger, 2008] Weinberger K. Q., Slaney M. & Van Zwol R. (2008). "Resolving Tag Ambiguity". In ACM Multimedia. pp. 111-120.

[Wu 1994] Wu Z. & Palmer M. (1994). " Verb semantics and lexical selection". In the 32 nd Annual Meeting of the Association for Computational Linguistics. , New Mexico, USA. pp. 133-138

[Zacklad, 2007] Zacklad M. (2007). "Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI) ". congrès annuel de l'Association Canadienne des Sciences de l'Information CAIS/ACSI. Partage de l'information dans un monde fragmenté : Franchir les frontières, Montréal. Disponible sur : http ://www.cais-acsi.ca/proceedings/2007/zacklad 2007.pdf

[Zhang, 2009] Zhang D., mao r., li W. (2009). "The recurrence dynamics of social tagging ". In the 18th international conference on World wide Web WWW 09. Madrid, Spain. pp. 1205-1206