République Algérienne Démocratique et Populaire

Ministère de l'Enseignement supérieur et de la recherche scientifique

Université SAAD DAHLEB De Blida

Faculté des Sciences

Département d'Informatique

Mémoire de fin d'étude

Pour l'obtention du diplôme de

MASTER

En Informatique

Option : Génie Logiciel

Thème

Extension et Réalisation de l'approche de filtrage de tags à base du profil utilisateur

Présenté par :

- **Foura Med Abdelatif**
- **Menacer Chahira**

Encadrés par :

Mme Kichou Saida : Attachée de recherche au CERIST

Suivis par :

Melle.Benblidia : Docteur

Année scolaire : 2012/2013

Remerciements

Nous remercions tout d'abord le bon dieu de nous avoir donné la santé, la volonté et le courage de pouvoir terminer ce projet.

Nous remercions chaleureusement notre encadreur madame « Kichou Saida » attachée de recherche au CERIST, de nous avoir donné l'occasion de travailler avec elle, nous la remercions d'avoir été tout le temps disponible et compréhensive, de nous avoir orienté, merci pour ses corrections, son aide et ses précieux conseils.

Nous remercions notre promotrice Mademoiselle « Benblidia », pour l'aide qu'elle nous a apporté.

Nous tenons à remercier le membre de jury d'avoir accepté de participer à notre soutenance.

Nous remercions infiniment nos familles et nos proches, d'avoir contribué à notre modeste travaille du mieux qu'ils puissent.

Nos remerciement les plus sincères !

Résumé

Le web 2.0 désigne une étape de l'évolution du web dont l'utilisateur et le partage d'information sont les clés de cette évolution, ce web qui est caractérisé par l'apparition de nouvelles fonctionnalités communautaires et collaboratives notamment le tagging qui favorise l'interaction entre les utilisateurs en leurs permettant d'attribuer leurs opinions et prestations à des ressources partagées (document, vidéo, image…) sous forme de mots clés (tags).

Une ressource est représentée par ses tags les plus populaires, cependant un tag populaire ne décrit pas toujours la ressource à laquelle il est associé, pour cette raison une approche de filtrage des tags à base du profil utilisateur précédemment proposée permettant l'intégration du profil utilisateur dans le calcul du poids des tags en se basant sur son expertise, ses centres d'intérêts ainsi qu'une auto-évaluation par l'utilisateur lui-même par rapport à la ressource.

En revanche cette approche représente certaines limites, ce qui nous empêche d'avoir des résultats exacts concernant le calcul du poids des tags, pour cela nous consacrons notre travail aux solutions permettant de pallier à ces limites ainsi nous pourrons avoir des tags reflétant au mieux la ressource à laquelle ils sont associés.

Mots-clés

Tagging collaboratif, tag, profil utilisateur, ontologie.

Abstract

Web 2.0 designates a step in the evolution of the web which the user and information sharing are the key of this development, this web which is characterized by the appearance of new community and collaboration features including tagging that promotes interaction between users by allowing them to allocate their opinions to shared resources (document, video, picture…) in the form of keyword (tag).

A resource is represented by its most popular tags, though a popular tag does not always describe the resource to which it is associated, for this reason, an approach of filter of tags based on user profile previously proposed allowing the integration of the user profile in the calculation of the weight of tags based on his expertise, interests center and a self-assessment by the user itself relative to the resource.

However this approach has some limits, which prevent us from having accurate results concerning calculate of the weight of tags, for this we dedicate our work to solutions to overcome these limitations, so we can have tags that reflects the resource to which they are associated.

Keywords

Collaborative tagging, tag, user profile, ontology.

ملخص

الويب 2.0 يشير إلى مرحلة في تطورشبكة الإنترنت، حيث أن المستخدم وتبادل المعلومات هي المفتاح لهذا التطور، هذا الويب الذي يتميز بظهور ميزات جديدة اجتماعية و تعاونية بما في ذلك التوسيم الذي يشجع التفاعل بين المستخدمين من خلال السماح لهم بتخصيص آراءهم بالنسبة للموارد المشتركة (فيديو، صورة...) بشكل كلمات مفتاحية.

المورد ممثل بالكلمات المفتاحية الأكثر استعمالا، ومع ذلك الكلمات المفتاحية الأكثر شعبية لا تصف دائما المورد الذي ترتبط به، لهذا السبب اقترح منهجا لتصفية الكلمات المفتاحية يستند على ملف المستخدم، الذي يسمح باستعمال معلومات المستخدم في حساب وزن الكلمات المفتاحية مستندا على خبرته، مجالات اهتماماته و تقييم ذاتي من قبل المستخدم نفسه بالنسبة إلى المورد.

ولكن هذا المنهج له حدود معينة التي تمنعنا من الحصول على نتائج دقيقة بالنسبة لحساب وزن الكلمات المفتاحية، لهذا السبب كرسنا عملنا في الحلول التي تمكننا من القضاء على هذه النواقص و هكذا يمكننا الحصول على كلمات مفتاحية تمثل بطريقة أفضل الموارد التي ترتبط بها.

الكلمات المفتاحية

التوسيم التعاوني، الكلمة المفتاحية، مواصفات المستخدم، علم الوجود.

Table des matières

Introduction générale

Première partie : Etat de l'art

Chapitre I: Le tagging collaborative

Chapitre II :

L'approche de filtrage des tags à base du profil utilisateur

Deuxième partie : Contribution

Chapitre III : Conception

Chapitre IV : Réalisation du système

Conclusion générale

Annexes

Liste des figures

Liste des tableaux

Liste des formules

Introduction Générale

Introduction

Considéré comme l'évolution naturelle du web actuel, le web 2.0 est un concept d'utilisation d'internet qui a pour but de valoriser l'utilisateur et ses relations avec les autres, en d'autres termes le web 2.0 est une plateforme d'échange entre les utilisateurs, cette évolution a conduit à l'émergence des systèmes du tagging collaboratif.

Le tagging collaboratif signifie la communication et la collaboration entre les utilisateurs, en leurs permettant de diffuser et de partager des ressources (document, vidéo…) en même temps de donner leurs avis et s'exprimer sur des contenus en ligne en attribuant librement des mots-clés, appelés aussi des tags.

Problématique

Le travail que nous présentons rentre dans le cadre d'améliorer les fonctionnalités des systèmes du tagging collaboratif, ces systèmes qui accordent aux utilisateurs de partager des ressources en lignes ainsi d'associer des mots-clés (tags) décrivant ces ressources, qui permettent par la suite l'exploration de ces ressources, il est donc primordial que les tags décrivent au mieux la ressource à laquelle ils sont associés.

Les systèmes du tagging collaboratif actuel considèrent les tags les plus populaires comme étant les tags représentant la ressource, or un tag populaire ne reflète pas forcément la ressource concernée, sachant que les utilisateurs ont tendance à répéter les mêmes tags, ce qui rend un tag populaire sans être vraiment représentatif de la ressource**[Wang, 10]**, à partir de cette hypothèse **[kichou,10]** a proposé une approche de filtrage des tags en se basant sur le profil utilisateur, cette approche consiste à calculer les poids des tags en prenant en considération le profil utilisateur (ses centres d'intérêt, son expertise et la confiance), afin de filtrer les meilleurs tags d'une ressource.

Partant du principe qu'un utilisateur est libre dans le choix de ses tags, nous réalisons que l'approche du filtrage de tag à base du profil utilisateur représente certaines limites, car un tag peut avoir plusieurs sens ce qui nous conduit a un problème d'ambigüité, comme il peut être écrit sous diverses formes ce que nous connaissons par la variation d'écriture.

L'idée donc de notre travail est d'exploiter cette approche et proposer des solutions afin de pallier à ses limites et de pouvoir sélectionner les descripteurs les plus exacts pour les ressources.

Organisation du mémoire

Notre mémoire est organisé comme suit :

Première partie : Etat de l'art

Dans cette partie nous allons étudier les systèmes du tagging collaboratif dans le premier chapitre, dans le deuxième chapitre nous aborderons l'approche de filtrage des tags à base du profil utilisateur.

- ***Chapitre I : tagging collaboratif***

Ce chapitre concerne les systèmes du tagging collaboratif donc nous commençons par décrire le tagging collaboratif ses structure, ses différents modèles et ses propriétés, nous parlerons aussi de la folkosonomie en précisant ses types, ensuite nous citant les limites des systèmes du tagging ainsi que les divers travaux de recherche concernant le tagging.

- ***Chapitre II : approche de filtrage des tags à base du profil utilisateur***

Dans ce chapitre nous décrivons le principe général de l'approche de filtrage des tags à base du profil utilisateur, nous parlerons aussi du modèle de profil utilisateur et de la pondération des tags, par la suite nous citons les points forts et les points faibles de cette approche et enfin nous présentons quelques solutions permettant de remédier aux problèmes de l'ambigüité et de variation d'écriture.

Deuxième partie : Contribution

Cette partie de notre mémoire contient deux chapitres :

- ***Chapitre III : conception du système***

Le chapitre III est le chapitre dans le quel nous définirons nos motivations ainsi que nos choix des solutions à réaliser en précisant les différentes étapes à suivre pour la conception de notre système et en l'illustrant par des exemples.

- ***Chapitre IV : réalisation du système***

Le dernier chapitre de notre mémoire concerne l'implémentation du système à réaliser, nous allons tout d'abord définir les outils et l'environnement de développement, ensuite nous passons à la présentation de l'application et ses différentes fonctionnalités.

Enfin nous terminerons notre mémoire par une conclusion générale résumant le travail que nous avons effectué ainsi que les perspectives possibles pour l'amélioration de notre travail.

Première Partie
Etat de L’art

CHAPITRE I : Tagging collaboratif

1. Introduction

L'organisation efficace de l'information est apparue comme une difficulté majeure dès le début de l'apparition d'Internet.

Aujourd'hui, les contenus disponibles sur le Net augmentent notamment sous l'effet de la participation de plus en plus active des utilisateurs qui deviennent créateurs de textes, d'images, de vidéos…etc qu'ils partagent sur la toile. Cette tendance se manifeste par la multiplication des blogs et autres espaces personnels. Ces nouvelles pratiques ont été accompagnées par l'apparition de nouveaux modes de navigation et d'organisation des contenus, dont le tagging collaboratif en ligne. Des millions d'utilisateurs l'utilisent quotidiennement sur des sites tels que Flickr, Delicious, Technorati…

Le tagging collaboratif ne cesse de gagner de popularité sur le web, cette nouvelle génération du web fait de l'utilisateur un lecteur-rédacteur.

Dans ce chapitre nous allons commencer par définir le tagging collaboratif, en citant ses structures et ses différents modèles, ensuite nous définirons la folksonomie et ses différents types, nous aborderons par la suite les études concernant les systèmes de tagging collaboratif, et enfin nous terminerons par citer les limites du tagging collaboratif.

2. Définitions

2.1. Tagging collaboratif

Le tagging collaboratif est décrit comme un processus pour de nombreux utilisateurs qui consiste à assigner des métadonnées sous forme de mot-clé à des contenus en ligne (vidéo, image, document .....) dans un environnement multi utilisateurs. **[Golder et Huberman, 06]**

2.2. Système du tagging collaboratif

Le système du tagging collaboratif est constitué des trois éléments suivant (fig1) :

- **L'utilisateur :** un internaute qui a la possibilité de partager ou annoter une ou plusieurs ressources.
- **Un tag :** un mot clé choisi par l'utilisateur pour décrire une ressource.

- **Une ressource :** un document de type : vidéo, image, texte... partagée par l'utilisateur.

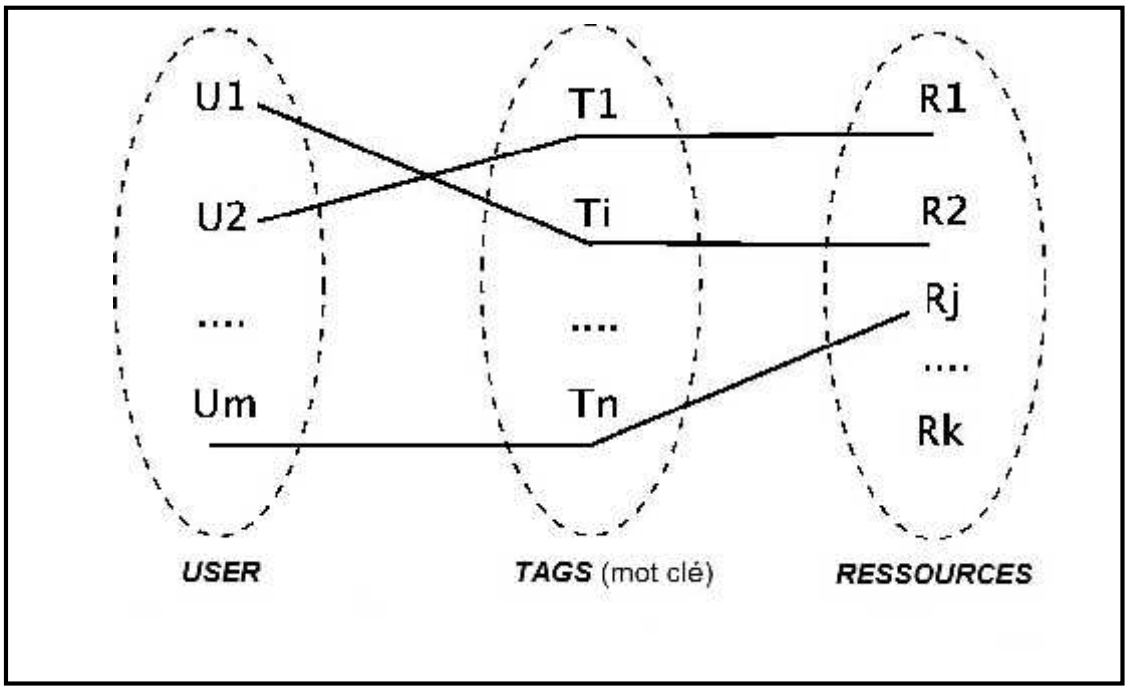


Fig 1: Le système de tagging collaboratif [Marlow, 06]

2.3. Tag

Le tag (étiquette) est un ou plusieurs mots clé choisi et attribué librement par des utilisateurs pour décrire une ressource partagée sur le web **[Guy et Tonkin, 06]**.

Pour pouvoir attribuer des tags, il faut faire partie d'une plateforme qui propose ce service (généralement gratuitement). Cela implique de créer un compte utilisateur (identifiant et mot de passe) sur un site Internet tel que Flickr ou Technorati.

Lorsqu'un contenu donné (vidéo, image.....) est étiqueté, on dispose de l'ensemble des tags proposés par les utilisateurs ayant accédés à ce contenu. Pour chaque tag attaché à ce contenu, on connaît en outre : la date de création du tag, le nombre de fois où il a été proposé pour ce même contenu, et le nom des utilisateurs (ou pseudonyme) qui l'ont choisi pour décrire ce contenu **[Roxin et Bernard, 07]**.

Donc la fonction principale des tags est d'aider les utilisateurs à mieux organiser et retrouver des contenus en ligne.

2.3.1. Classification des tags

D'après **[Wetzker et al, 10]** les tags sont classifiés selon leurs :

- Signification : les tags peuvent avoir un aspect :

- **Personnel :** c'est-à-dire qu'ils ont un sens que pour les personnes qu'ils l'ont utilisé et ne signifient rien pour les autres utilisateurs.
- **Collectif :** c'est-à-dire qu'ils ont le même sens pour tout les utilisateurs.

- la fréquence d'apparition dans la ressource : les tags peuvent être :

- **Significatifs pour la ressource:** l'étiquette se trouve dans le document et le représente.
- **Non significatifs:** l'étiquette peut se trouver ou pas dans le document, mais elle n'est pas représentative pour la ressource.

- Popularité : Les tags populaires sont les tags les plus utilisés par l'ensemble des utilisateurs ou les tags les plus connus par ces derniers, ces tags ne sont pas forcément précis ou représentatifs de la ressource et souvent se sont des termes génériques.

2.3.2. Type des tags

Selon le site de tagging collaboratif *del.icio.us* **[Golder et Huberman, 06]** citent ces différents types de tag :

- **Subjectifs :** tags qui explicitent une opinion ou une émotion, par exemple : effrayant, froid.
- **Descriptifs :** lorsque le tag décrit la ressource d'une façon globale.
- **Populaires :** se sont les tags les plus utilisés par les utilisateurs.
- **Evaluateurs :** les tags qui contiennent des chiffres importants, par exemple : windows8.
- **Self référence :** les tags commençant par « my » (my_comments) concernant la langue anglaise, ils identifient une relation entre l'utilisateur et la ressource.

3. Structure de tagging collaboratif

Il existe plusieurs structures des systèmes du tagging collaboratif, mais la principale structure est tripartie, cependant il y a la structure tripartie avec des liens et quadripartite.

3.1. Structure tripartite

[Halpin et al, 07] décrit le système de tagging collaboratif par un réseau tripartite, les nœuds étant des utilisateurs, des documents et des tags comme le montre la figure suivante :

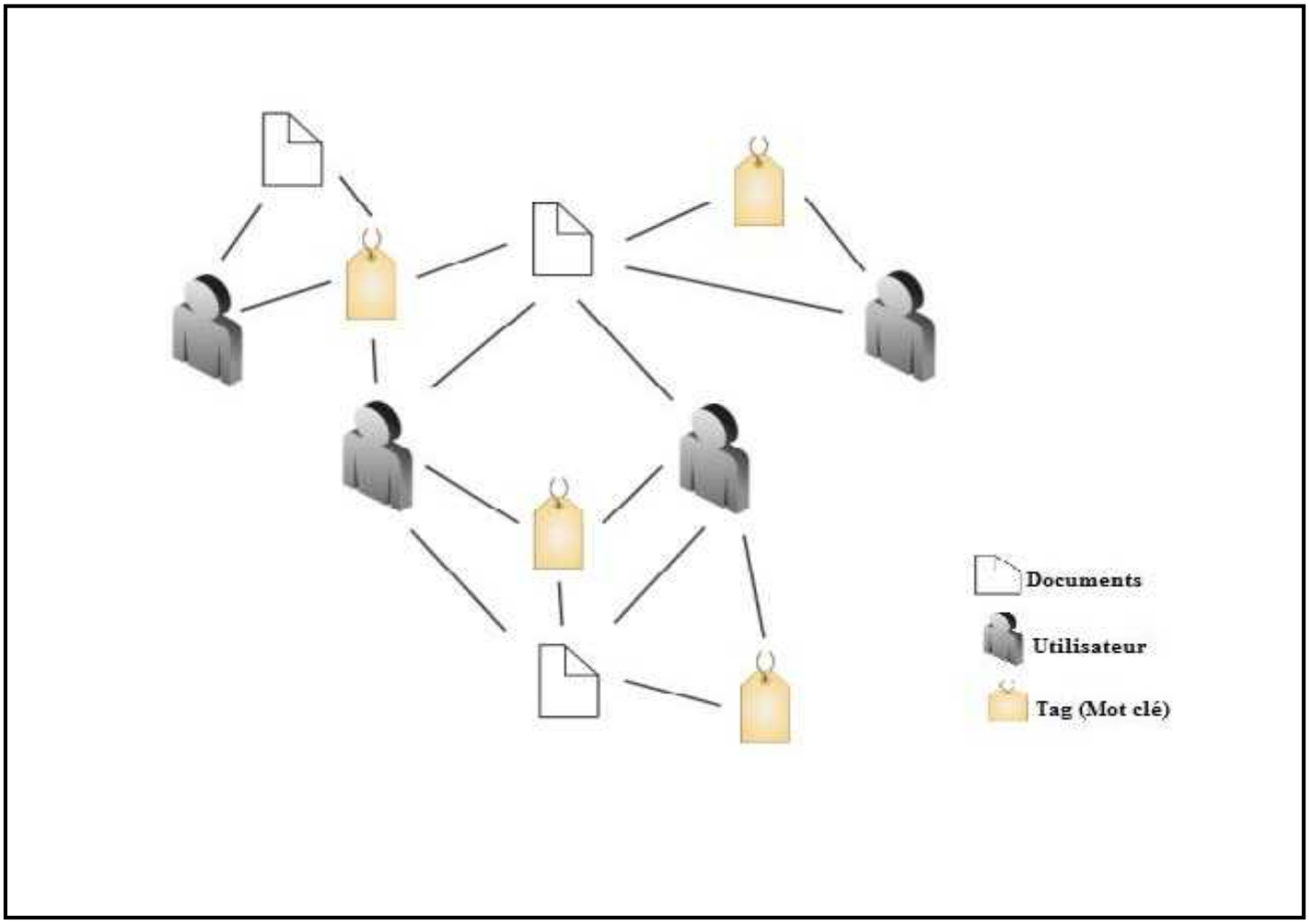


Fig 2 : Structure tripartite d'un système de tagging collaboratif [Halpin et al, 07]

- **Autour des documents**

Concentrons-nous premièrement sur les documents. Nous voyons qu'autour de chaque document, il y a un certain nombre d'utilisateurs et un certain nombre de tag :

1. Le voisinage d'utilisateurs nous fournit les personnes qui s'intéressent à ce document.
2. Le voisinage des tags nous fournit une indication de son contenu (exp : java, windows .....).

- **Autour des mots clés (tag)**

Examinons deuxièmement les mots clés, le voisinage d'un mot clé est constitué :

1. De tous les documents pour lesquels au moins un utilisateur a jugé que le mot clé s'appliquait, c'est l'équivalent des résultats d'une recherche par mot clé.
2. De tous les utilisateurs qui ont utilisé ce mot clé, c'est-à-dire que c'est un groupe formé de gens qui ont un intérêt en commun.

- **Autour des utilisateurs**

Finalement, observons le voisinage des utilisateurs. Autour d'un utilisateur on retrouve ses mots clés, qui reflètent ses intérêts, ainsi qu'un ensemble de documents, qui peuvent être vu comme une sorte de bibliothèque personnelle également indicative de ses intérêts.

3.2. Structure tripartite avec liens inter-utilisateurs et inter-ressources

Dans une structure tripartite du tagging collaboratif, des liens peuvent être présent entre les ressources, ou entre des utilisateurs **[Marlow, 06]**. Nous pouvons voir ceci dans le modèle conceptuel suivant :

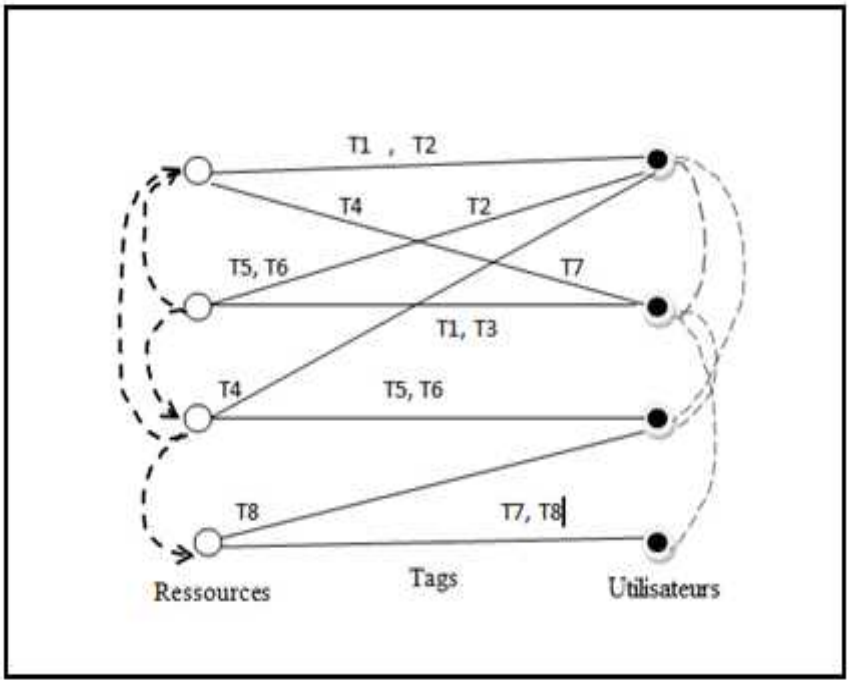


Fig 3 : Modèle conceptuel d'un système de tagging avec liens [Marlow, 06]

Dans ce modèle conceptuel les tags sont représentés sous formes d'arrêtes reliant les ressources aux utilisateurs, les liens entre utilisateurs ou entre ressources sont représentés en pointillés.

Même si ces liens n'existent pas directement, nous pouvons dire qu'il existe toujours des relations indirectes entre ressources à travers les tags des utilisateurs (utilisateurs communs), et de même pour les utilisateurs à travers les ressources qu'ils tagguent (ressources communes entre un ensemble d'utilisateurs).

3.3. Structure quadripartite :

[Gruber, 07] a proposé une extension de la structure tripartie en ajoutant le concept *source* (signification) faisant référence à la signification du tag, par exemple : le tag « Apple », peut signifier la marque Apple **Tagging1** (user1, photo, Apple, Apple_computers) comme il peut s'agir d'un fruit **Tagging2** (user2, photo, Apple, fruit).

La structure d'une action du tagging est dans ce cas un quadruplet : tagging (utilisateur, ressources, tag, source). Ce modèle est utilisé dans le tagging guidé par ontologie.

La figure suivante représente la structure quadripartite d'un système du tagging :

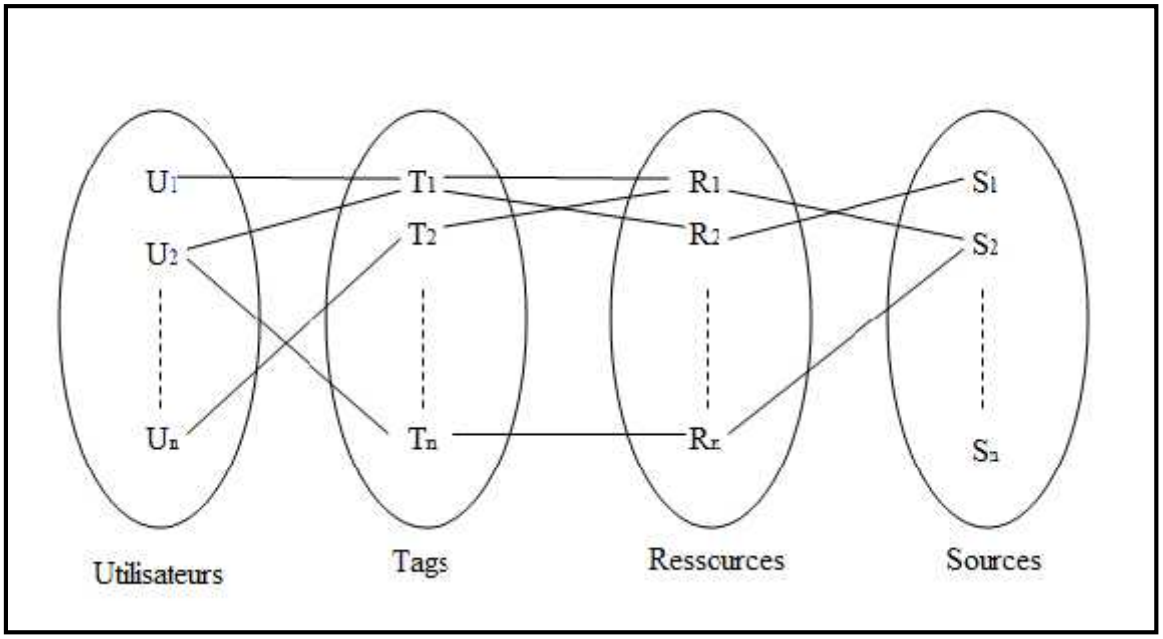


Fig 4 : La structure quadripartite d'un ensemble d'actions du tagging collaboratif [Gurber, 07]

4. Les modèles du tagging collaboratif

Selon **[Hedstrom, 05]** les différents modèles du tagging collaboratif peuvent être caractérisés par un modèle descriptif ou prédictif.

4.1. Modèle descriptif

Le modèle descriptif du tagging collaboratif décrit le comportement des utilisateurs devant les contenus taggués.

[Cattuto et al, 07] Le modèle stochastique MBYS (The Memory-Based Yule-Simon) explique l'affectation des tags à une ressource. A chaque pas de temps une nouvelle étiquette est assignée avec une probabilité ***p*** et une étiquette est copié (imiter) par un tag existant avec probabilité ***1-p***. Lors de la copie, la probabilité de sélectionner un tag dans l'ensemble des tags existants décroît avec le temps. Et cette décroissance suit une distribution en loi de puissance.

Donc ils ont constaté qu'avec ce modèle les tags récemment utilisées sont supposées avoir une probabilité plus élevée d'être imités et donc devrait de plus en plus dominer le tag-vocabulaire.

4.2. Le modèle prédictif

[Fu, 08] a adopté le processus d'imitation sémantique dans le modèle prédictif. Il se concentre sur l'interprétation sémantique des tags. Comme proposé, lors de la recherche d'informations, les utilisateurs perçoivent les tags comme des indices importants et les traitent de manière intensive pour déduire des contenues dans les ressources. Ensuite, les utilisateurs combinent les thèmes du document et celles extraites des tags pour décrire la ressource avec des tags propres. En d'autres termes, les tags invoquent une sémantique d'interprétation, ce qui affecte la compréhension d'une ressource.

5. Les propriétés des systèmes de tagging collaboratif

Le système de tagging collaboratif offre différentes propriétés aux utilisateurs, entre autres nous citons :

- La liberté de choix des mots-clés (tags).
- La possibilité de tagguer ses propres ressources, ainsi que les ressources des autres utilisateurs.
- La possibilité d'associer plusieurs mots-clés pour la même ressource.
- Un même tag peut être associé à différentes ressources par différents utilisateurs.
- Une ressource est tagguée pas plusieurs utilisateurs.
- L'utilisateur doit s'identifier.

6. La folksonomie

Le mot folksonomie est apparu sur le web pour décrire une expérience de classement de ressources effectué librement et spontanément par des utilisateurs du réseau. Ce néologisme de l'anglais est né de la contraction des mots « folks » (les utilisateurs) avec « taxonomy » (la classification d'éléments).

Plus largement, la folksonomie désigne les actions de classification, annotation et indexation de tous types de documents sur le web sans aucune règle. L'objectif est de bénéficier des efforts de tous pour trouver plus facilement et plus vite des ressources plus adéquates à notre besoin. [**Vander Wal ,07]**

La folksonomie, ce phénomène relativement récent qui commence à susciter de nombreuses recherches en informatique permet de :

- **Améliorer la recherche :** la folksonomie permet de classer et indexer les ressources.
- **La facilité d'usage :** car elle est peu couteuse.
- **La folksonomie est dynamique** : mise à jour automatique et en permanence puisque le tagging est continu.
- **Un moyen de veille et de mesure :** la folksonomie permet la traque de termes précis (*tracking*). Elle est également utilisée pour mesurer la popularité de tags, de blogs.
- De plus la rapidité de la diffusion de l'information.

6.1. Les types de la folksonomie

Comme l'explique [**Vander Wal, 05]**, nous pouvons distinguer deux typologies de folksonomie, chacun associé à des propriétés spécifiques et des suggestions d'utilisation:

- **Folksonomie générale (broad folksonomy)**

Une folksonomie générale (comme celle de Del.icio.us) est le résultat de l'étiquetage de nombreux utilisateurs de la même ressource. Chaque utilisateur peut marquer l'objet d'une façon différente suivant son propre modèle, vocabulaire et langage.

Ce type de folksonomies s'appuie sur des réseaux sociaux ne fait pas que classer de l'information et la partager, mais également il met en relation des utilisateurs qui partagent les mêmes centres d'intérêts.

- **Folksonomie étroite (narrow folksonomy)**

Une folksonomie étroit est le résultat d'un plus petit nombre d'utilisateurs marquant (en utilisant un ou plusieurs tags) des ressources pour une récupération ultérieure personnel ou à leur propre convenance.

D'un autre terme c'est l'indexation par l'utilisateur de ses propres ressources.

6.2. Les nuages de mots

Les représentations visuelles des folksonomies permettent aux utilisateurs de les exploiter pour leurs recherches et leurs activités d'étiquetage. Il en existe plusieurs : diagrammes, réseaux sémantiques (où chaque nœud représente un *tag*)...

Mais la représentation la plus commune sur les sites Web est le nuage de mots. Son succès provient probablement de sa facilité d'utilisation et de sa capacité à fournir de manière simple un assez grand nombre d'informations.

Un nuage de tags est une présentation visuelle des tags utilisés pour décrire les ressources d'un site Web particulier, ou des contenus extérieurs mais indexés sur ce site **[Sinclair et Cardew-Hall, 08]**.

Flickr est le premier site à avoir implémenté les nuages de mots comme une forme d'affichage des tags, par la suite les nuages de mots ont été utilisé sur d'autres sites tel que Delicious, Technorati … (quelques sites de tagging sont représentés dans l'Annexe A).

Selon **[Roxin et Bernard, 07]** pour faciliter la recherche de contenus, un nuage de mots peut être construit selon divers paramètres :

- **L'ordre des tags :** il peut être alphabétique ou en fonction de la popularité des mots.
- **La couleur de tags :** elle peut indiquer l'origine des tags (l'utilisateur lui-même, les autres, les utilisateurs les plus populaires, etc.).
- **Le contraste des tags :** les plus foncés sont les plus récents.
- **La taille des tags :** les plus populaires sont affichés en taille plus grande. Deux standards sont utilisés :

- Soit la taille représente le nombre de fois que le tag en question a été attribué à un contenu donné.
- Soit elle représente le nombre de contenus qui ont été étiquetés avec chaque tag.

Les nuages de mots sont généralement interactifs c'est à dire chaque tag affiché est un lien vers une page de résultats qui contient une liste des contenus qui ont été indexés avec le mot en question (fig 5).

Fig 5: les nuages de mots

7. Les limites des systèmes du tagging

Selon **[Guy et al, 06], [Sen et al, 06]** et **[Passant, 07]** la liberté du choix des tags provoque certaines limites dans le système de tagging, parmi ces limites :

- **L'ambigüité :** tient à l'existence de l'homonyme (le même mot mais sens différent) Exemple : le mot *avocat* qui peut signifier soit une profession ou bien un fruit.
- **Problème de variation d'écriture :** c'est lorsque un même tag se présente de différente formes par exemple le mot *email* et *e-mail*.
- **Problème typographique :** c'est le cas d'utilisation des termes incompréhensibles en dehors d'un sous-groupe de référence ou lorsqu'un utilisateur emploie la forme au pluriel et un autre la forme singulière d'un même tag par exemple : *programme* et *programmes* Dans ce cas le système sera incapable de faire le rapprochement entre ces différents tags.
- **Problème du multilinguisme :** selon **[Wetzker et al, 10]** les utilisateurs parlant des langues différentes impliquent l'utilisation des tags différents (*music* en anglais *musique* en français).
- **Problème du vocabulaire :** fautes de grammaires et fautes de typage. **[Furnas et al, 87]**
- **info-pollution :** des utilisateurs malveillant peuvent spammer (mettre des contenu non sollicité) le système.

8. Les règles d'indexation collaborative

L'indexation collaborative, née du processus de tagging collaboratif est une classification faite par les utilisateurs.

Afin de diminuer les problèmes que nous avons soulevés concernant l'utilisation de tagging collaboratif, **[Deuff, 06]** a dicté les règles suivantes pour une bonne indexation :

- L'utilisateur doit penser collectivement : les tags sont certes personnels mais peuvent être utilisés par d'autres utilisateurs.
- Employer le pluriel pour définir des catégories : le pluriel est plus approprié car les catégories peuvent contenir des variations.
- Ne pas employer la majuscule.
- Inclure des synonymes afin d'éviter des confusions.
- Définir un groupe de mots grâce à l'undrescore.(_)
- Observer et utiliser les conventions d'indexations des réseaux sociaux utilisés.

9. Etude des systèmes du tagging collaboratif

Ces dernières années, le tagging collaboratif s'est imposé au sein du Web comme le principal moyen de classification et d'organisation de données.

Cette popularité a créé de nouveaux défis, les utilisateurs participent massivement à la production de l'information et l'organisation de celle-ci.

Ces nouvelles fonctionnalités sont l'objet de plusieurs travaux recherche.

Dans cette partie nous allons classifier les différentes études sur les systèmes du tagging collaboratif.

9.1. Etude de la dynamique du tagging

L'étude de la dynamique des systèmes de tagging collaboratif est un domaine de recherche actif ces dernières années.

De nombreux facteurs tels que : la fréquence d'utilisation des tags, la variation des nombres de tags utilisés par les utilisateurs, les types de tags … sont pris en considération par les auteurs pour cette étude.

[Golder et Huberman, 05] sont les premiers à avoir analysé la dynamique des systèmes de tagging collaboratif, à l'aide des données de *del.ici.ous* ils ont noté un certain nombre de modèles de la dynamique de tagging.

Après leurs analyse ils ont remarqué que la majorité des sites atteignent leur pic de popularité et la fréquence la plus élevée de marquage dure un certains temps (environs 10 jours après avoir été enregistré sur *del.icio.us*) et ils ont observé que les utilisateurs présentent une grande variété dans leurs tags, et que les tags eux mêmes varient du point de vue de fréquence d'utilisation.

Cependant il existe trois hypothèses en ce qui concerne la dynamique de systèmes de tagging :

- **La convergence des tags :** de sorte qu'après un certain temps le nombre de tags à une ressource donnée reste stable et un même tag devient le plus courant et majoritaire pour cette ressource.
- **Divergence des tags** : serait que le nombre des tags ne deviendra pas un groupe de mots stable, et les tags changent continuellement.
- **Périodicité des tags :** c'est-à-dire un groupe de tagueurs assignent un ensemble de tags qui sont de la même catégorie, et après une période un autre groupe utilise un ensemble divergent et perturbe ainsi le jeu initial des tags. Ce processus peut se répéter et ainsi conduire à une convergence après une période d'instabilité.

Selon **[Marlow, 06]**, il existe des facteurs qui peuvent influencer la dynamique de système du tagging, ces facteurs appartiennent a deux catégories : conception de ces systèmes et motivation des utilisateurs.

La conception des systèmes comporte les facteurs suivant :

- **Les droits de tagging :** le système est soit en *self-tagging*, l'utilisateur ne peut tagguer que ses propres ressources, ou bien *free_for_all-tagging* l'utilisateur à le droit de tagguer d'autres ressources.
- **Type du tagging :** le tagging peut être soit *blind tagging* c'est-à-dire que l'utilisateur tagguant une ressource ne voit pas les tags attribués par les autres a cette même ressource, ou bien le tagging peut être *suggestive tagging* c'est-à-dire le système suggère aux utilisateurs une liste de tags.

- **Type d'objet à tagguer :** le type d'objet à tagguer peut être décisif pour le choix de tags, les ressources à tagguer peuvent être de type : vidéo, image, audio …

Parmi les motivations des utilisateurs nous citons :

- **Partage :** l'utilisateur peut partager ses ressources pour une ou plusieurs personnes.
- **Utilisation futur :** l'utilisateur taggue une ressource pour pouvoir l'utiliser ultérieurement.
- **Attirer l'attention :** les tags permettent de donner au système un style attirant et visible.
- **Jeux et compétition :** c'est le cas des jeux sur le tagging.

9.2. Travaux sur la proposition d'algorithme de suggestion de tags et d'utilisateurs

Un algorithme de suggestion de tags et d'utilisateurs a été développé dans le but d'orienter le choix de l'utilisateur pour une meilleur sélection de tags pour une ressource donnée, comme il permet de suggérer a l'utilisateur une liste d'autres utilisateurs qui partagent les mêmes centre d'intérêts que lui et finalement recommander une liste de sites pour un utilisateur en se basant sur ses informations personnelles (profil).

La suggestion des tags est une forme de tagging semi-automatique, qui toute fois permet aux utilisateurs de choisir entre la liste des tags suggérés ou introduire leur propre tag manuellement.

10. Conclusion

Le tagging collaboratif s'est récemment imposé dans le paysage du web social (Web 2.0) comme support à l'organisation de ressources partagées en permettant aux utilisateurs de catégoriser ces ressources, simplement en leur associant des mots clés.

Dans ce chapitre, nous avons défini le tagging collaboratif, ses structures et ses modèles, nous avons également décrit la folksonomie en indiquant ses types, nous avons cité les différentes règles d'indexation et par la suite nous avons abordé les travaux de recherche sur les systèmes de tagging collaboratif.

Actuellement les ressources sont représentées par leurs tags les plus populaires, mais comme les systèmes du tagging accordent aux utilisateurs la liberté du choix des tags, alors un

tag populaire ne représente pas toujours le contenu qu'il décrive, pour cela une approche de filtrage des tags a été proposée et qui sera l'objet du chapitre suivant.

CHAPITRE II : L'approche de filtrage des tags à base du profil utilisateur

1. Introduction

Les sites de tagging collaboratif ont vu leur popularité croître ces dernières années en raison de la démocratisation du web et de l'augmentation exponentielle de la quantité de ressources disponibles et accessibles, de nombreux sites tels que Flickr,Technorati intègrent un système de tagging.

Le système du tagging collaboratif a pour but de préconiser à un utilisateur des ressources en lien avec ses goûts et ses attentes. L'objectif est à la fois de minimiser son temps passé à la recherche, mais aussi de lui suggérer des ressources pertinentes qu'il n'aurait pas spontanément consultées et ainsi accroître sa satisfaction globale.

Actuellement ces systèmes désignent des tags populaires qui sont généralement affichés sous forme de nuages ou de listes de tags. Une ressource est représentée par ses tags les plus populaires, classés par ordre décroissant, or un tag populaire n'est pas toujours représentatif du contenu qu'il décrive **[Wang, 10]**.

Pour remédier à ce problème une approche de filtrage a été proposée et celle-ci intègre le profil utilisateur lors du calcul du poids des tags associé à une ressource donnée **[Kichou, 10]**.

Dans ce chapitre nous allons présenter cette approche ainsi que ces différentes étapes, la modélisation du profil utilisateur, la pondération des tags, et la création du descripteur de ressource composé des tags jugés plus représentatifs, par la suite nous allons désigner les points forts et les points faibles de cette approche et enfin nous assignerons les diverses solutions proposées pour pallier à certaines limites.

2. Principe général de l'approche

L'objectif de cette approche est de trouver un autre critère pour classer les tags afin de décrire les ressources de la manière la plus précise et la plus exacte possible **[Kichou, 10]**.

Cette approche consiste à intégrer l'utilisateur dans le calcul du poids des tags associés à une ressource donnée.

La figure suivante illustre le principe général de l'approche :

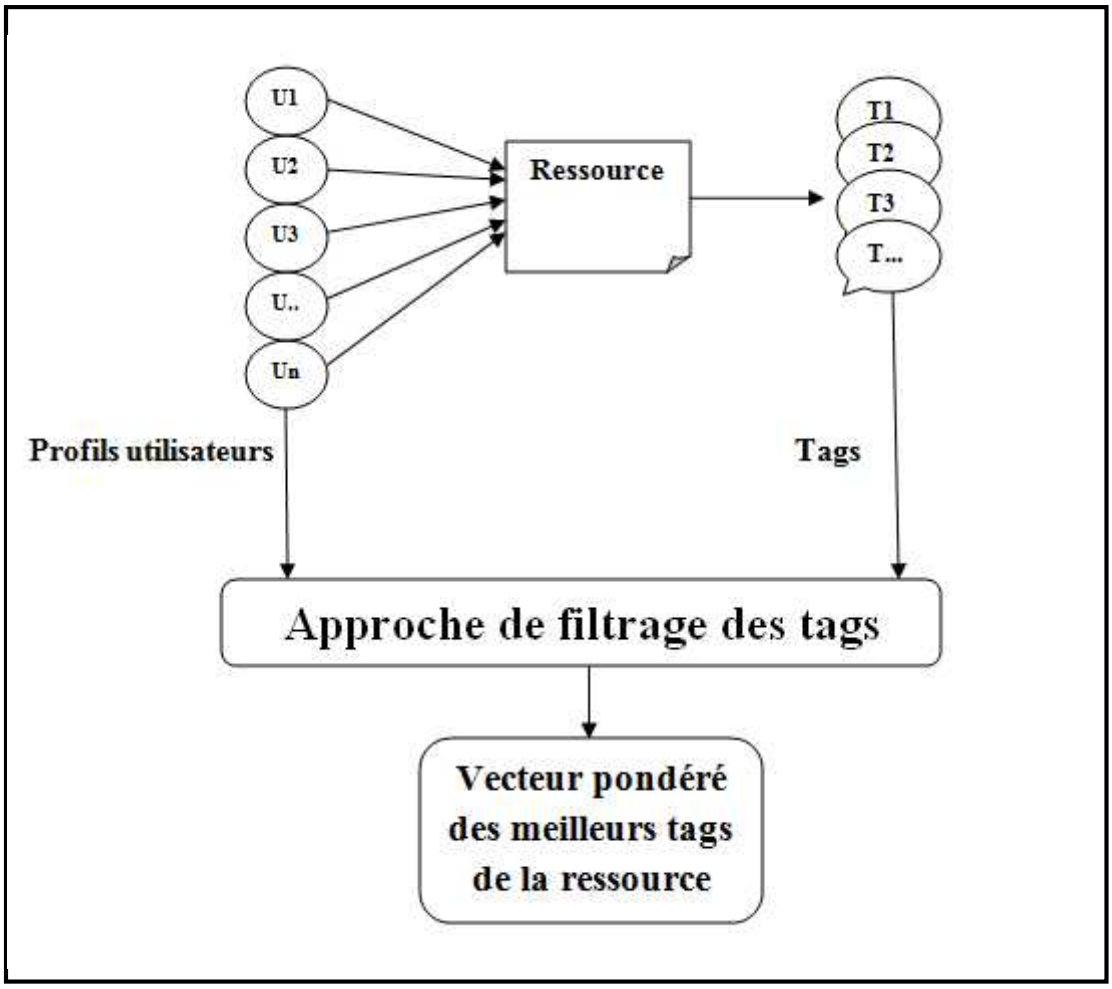


Fig 6 : Principe général de l'approche [Kichou, 10]

3. présentation de l'approche du filtrage des tags

L'approche se compose de trois principales étapes :

- Modélisation et construction du profil utilisateur.
- Pondération et classement des tags.
- Sélection des descripteurs des ressources.

La première étape consiste à créer et modéliser un profil en construisant ses dimensions, dans la deuxième étape la pondération des tags est affichée en prenant en considération les profils des utilisateurs, finalement la troisième étape est consacrée à la création de descripteur de la ressource composé des meilleurs tags (Filtrage).

Voici un schéma représentant l'approche du filtrage des tags :

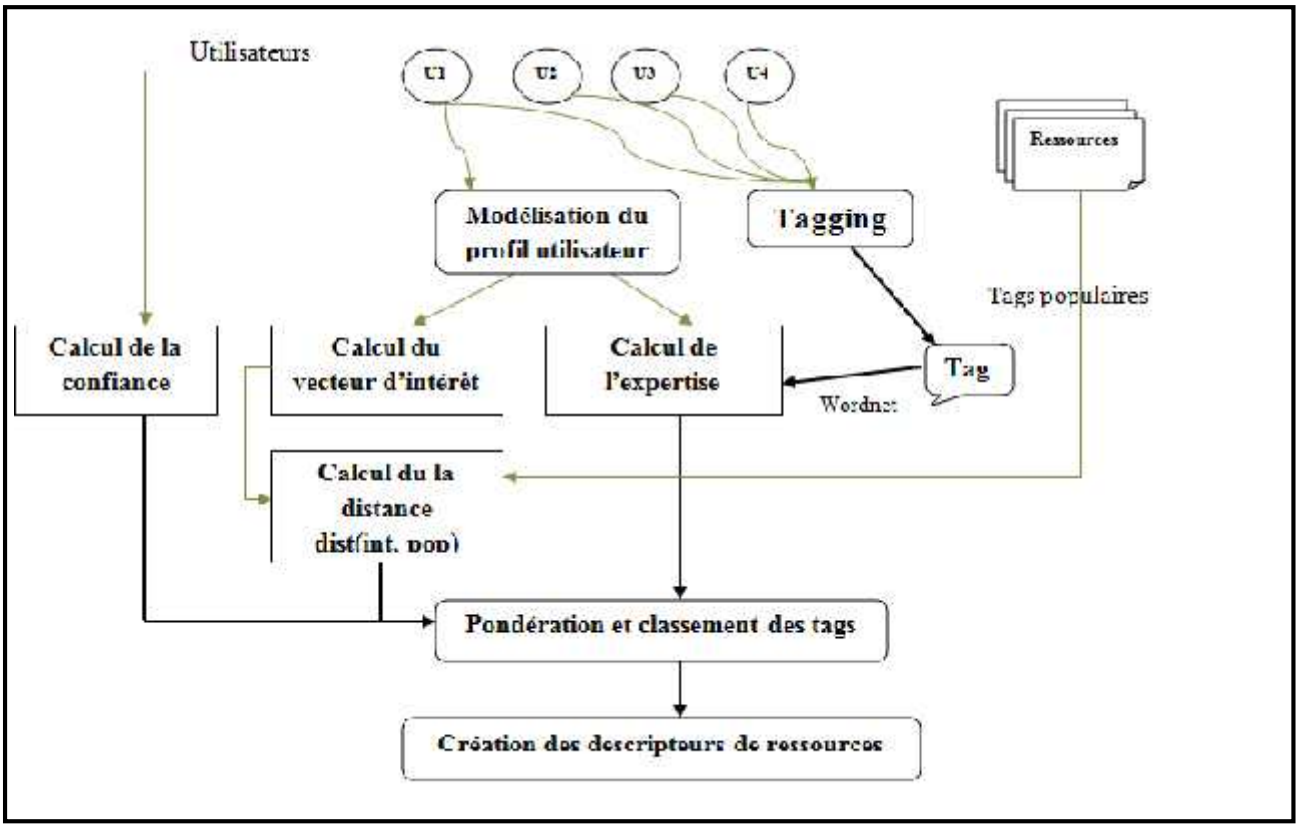


Fig 7 : Schéma global de l'approche

3.1. Le modèle du profil utilisateur

Afin d'intégrer le profil utilisateur dans le calcul du poids des tags, celui-ci doit être modélisé afin de contenir des informations reflétant ses activités dans le système. Nous allons présenter le modèle du profil utilisateur, tel qu'il est définit dans **[Kichou, 10]** ainsi que la démarche de sa construction.

3.1.1. Représentation du profil

Selon l'approche le profil utilisateur est multidimensionnel constitué de trois dimensions (fig8).

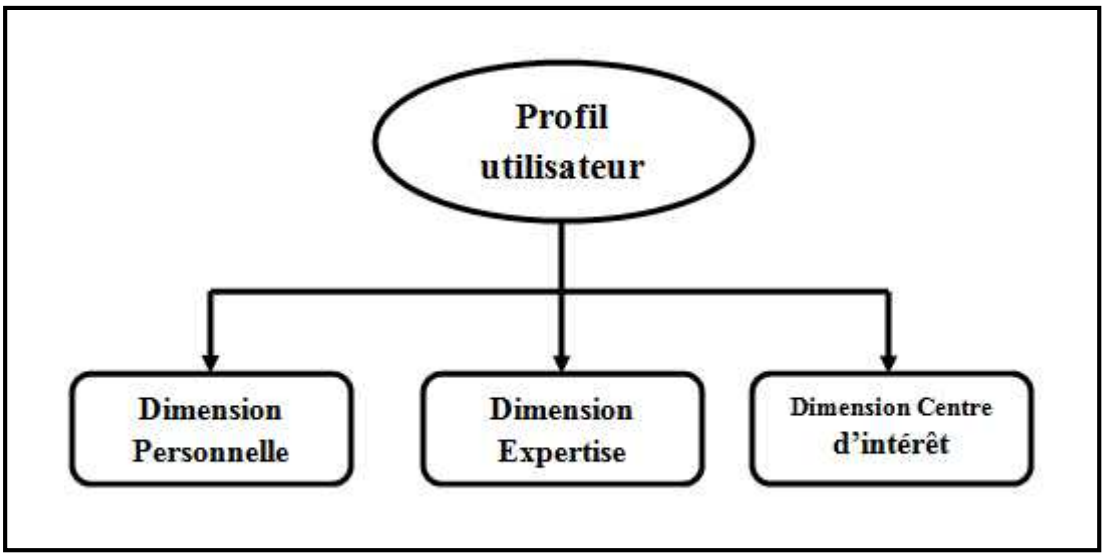


Fig 8 : Présentation du profil utilisateur

- **Dimension personnelle :** contient les informations personnelles de l'utilisateur qui servent à l'identifier (identifiant, nom, prénom, mot de passe …) ces informations sont introduite par l'utilisateur lors de son inscription.
- **Dimension expertise :** cette dimension est le degré de maitrise de l'utilisateur du domaine des ressources taguées.

  Les utilisateurs ayant une parfaite maitrise dans un domaine, ont tendance à utiliser des termes spécifiques de ce domaine.

  Par exemple : un utilisateur qui associe les tags « c++ », « php » est considéré plus expert qu'un utilisateur associant les tags « language », « programmation » dans le domaine informatique.
- **Dimension centre d'intérêt :** Les centres d'intérêt de l'utilisateur, constituent la troisième dimension du profil utilisateur traduisant ses préférences en se basant sur les tags qu'il associe, cette dimension est précisément construite selon l'historique de ses tags. La dimension centre d'intérêt *Int (ui)* est représentée sous forme de vecteur de tags pondérés, ce vecteur est construit en utilisant deux approches de construction de profil, l'approche naïve **[Cayzer et Michlmayr, 09]** et l'approche par cooccurrence **[Wasserman, 94]**.

  Int (u_i)= {(T_1, W_1), (T_2, W_2), (T_3, W_3)….. (T_n, W_n)}, tel que T représente le tag et W représente le nombre de fois que l'utilisateur a associé ce tag.

3.1.2. La construction du profil utilisateur

La construction du profil utilisateur est la construction des dimensions $Int(U_i)$ et $Exp(U_i)$ **[Kichou, 10]**.

- **Construction de la dimension centre d'intérêt**

Il existe plusieurs approches de construction d'intérêt des utilisateurs, cependant dans l'approche de filtrage deux approches les plus utilisées ont été combiné, il s'agit de l'approche naïve et l'approche par cooccurrence.

- *L'approche hybride (naïve / par cooccurrence)*

La combinaison de l'approches naïve et par cooccurrence élimine le problème de négligence des ressources à tag unique, mais aussi permet de pondérer les tags chose qui n'est pas permise avec l'approche par cooccurrence.

Le résultat de la combinaison est un graphe de nœuds (tags) et d'arcs pondérés, les nœuds pondérés appartenant aux arcs ayant le plus grands poids composent le vecteur d'intérêt.

Le graphe suivant représente la combinaison entre les deux approches :

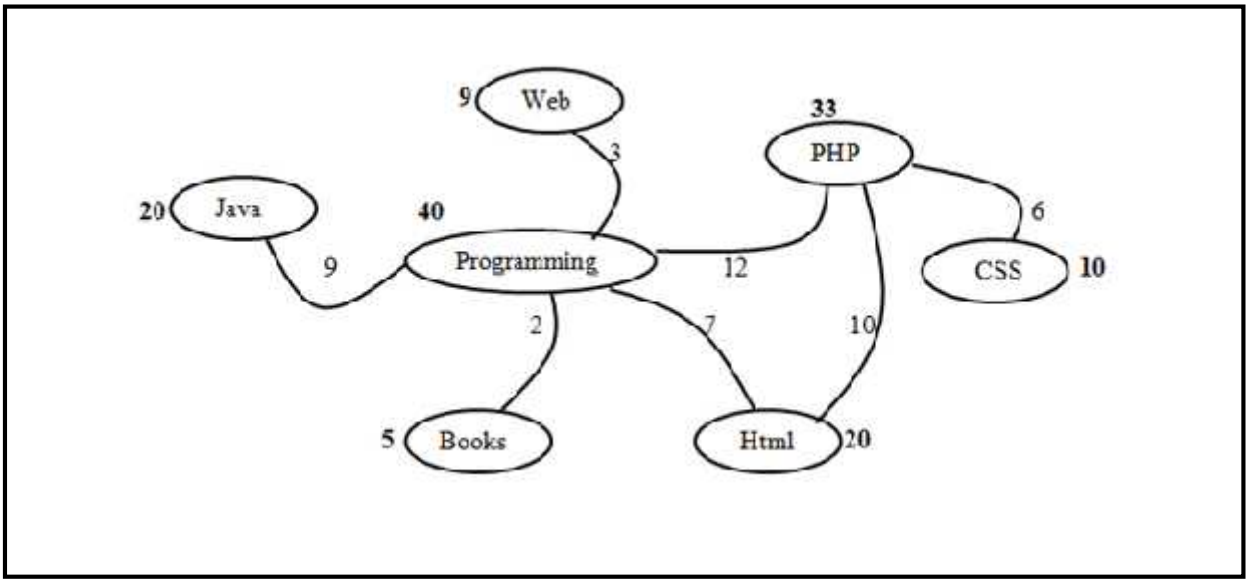


Fig 9 : Exemple d'un graphe construit avec la combinaison de l'approche naïve et l'approche par cooccurrence

Cette figure (Fig 9) illustre le graphe du profil d'un utilisateur avec les poids de cooccurrence des tags ainsi que la popularité de chaque tag.

Le tableau suivant est obtenu en choisissant les couples de tags participant aux arcs fortement pondérés :

Nombre de cooccurrence	Les tags et leurs poids
12	Programming(40), PHP(33)
10	PHP(33), Html(20)

Tableau 1 : top 2 des combinaisons de tags classées par l'approche hybride

- *L'algorithme Add-A-Tag adapté à l'approche hybride*

[Cayzer, 09] a proposé cet algorithme dans le but de mettre en œuvre une nouvelle approche de construction de profil, mais dans ce cas l'algorithme est adapté afin de réaliser l'approche hybride.

Soit un utilisateur **u** taggant un ensemble de ressources avec l'ensemble des tags T= {T_1, T_2, T_3……,T_n}, le graphe du profil utilisateur $\mathbf{G_u}$(V,E) où V={V_1, V_2, V_3,……,V_n} l'ensemble des nœuds et E={E_1, E_2, E_3……,E_n}l'ensemble des arcs.

Etape 1 : Mise à jour du graphe

Les **n** nouveaux tags introduit par un utilisateur **u** pour une ressource donnée sont ajoutés au graphe. Pour toute combinaison t_i t_j où i, j ∈ à 1, 2, 3….n et i<j la procédure suivante est exécutée :

1- Pour chaque tag t_x avec x ∈ i, j ajouter au graphe le nœud correspondant V_x si celui-ci n'existe pas ;
2- Si le nœud n'existe pas, créer un arc de poids égal à 1 entre le nœud V_i et le nœud V_j ;
3- Si le nœud existe déjà, incrémenter de 1 le poids de l'arc entre V_i et V_j ;
4- Affecter pour chaque nœuds du graphe son poids (sa popularité) ;

Etape 2 : Extraction du profil

1- Créer un sous ensemble Es de E, ordonné avec un ordre décroissant des poids des arcs ;
2- Choisir le top k des éléments de Es avec k un entier non nul ;
3- Ajouter au profil les tags correspondants aux arcs élus et leurs poids (popularités).

La taille du profil est déterminée par la valeur du paramètre k. c'est un vecteur de termes (tags) pondérés.

- **Construction de la dimension expertise**

Un utilisateur expert dans un domaine a tendance à associer des termes spécifique aux ressources qu'il taggue concernant ce domaine. Dans l'approche du filtrage une ontologie est utilisée permettant de situer les niveaux des utilisateurs (profondeur) dans l'hiérarchie de l'ontologie, plus le tag utilisé est profond plus l'utilisateur est expert.

WordNet est l'ontologie choisie par **[Kichou, 10]** afin de localiser les niveaux des tags utilisés pour le calcul de l'expertise.

Wordnet est une vaste base de données lexicale de la langue anglaise (noms, verbes, adjectifs et adverbes qui sont regroupés en ensembles de synonymes (Synset), les Synset sont reliés entre eux par des relations sémantique et chaque Synset exprime un concept. Les liens sémantiques à proprement parler ne relient alors pas les mots entre eux mais les Synset auxquels les mots sont affectés).

- *Les profondeurs des termes dans WordNet*

Le système de WordNet repose sur des ensembles de synonymes les Synset (Synonym set) qui contiennent un groupe de mots interchangeable.

La figure suivante montre le sens du mot 'science' en utilisant WordNet :

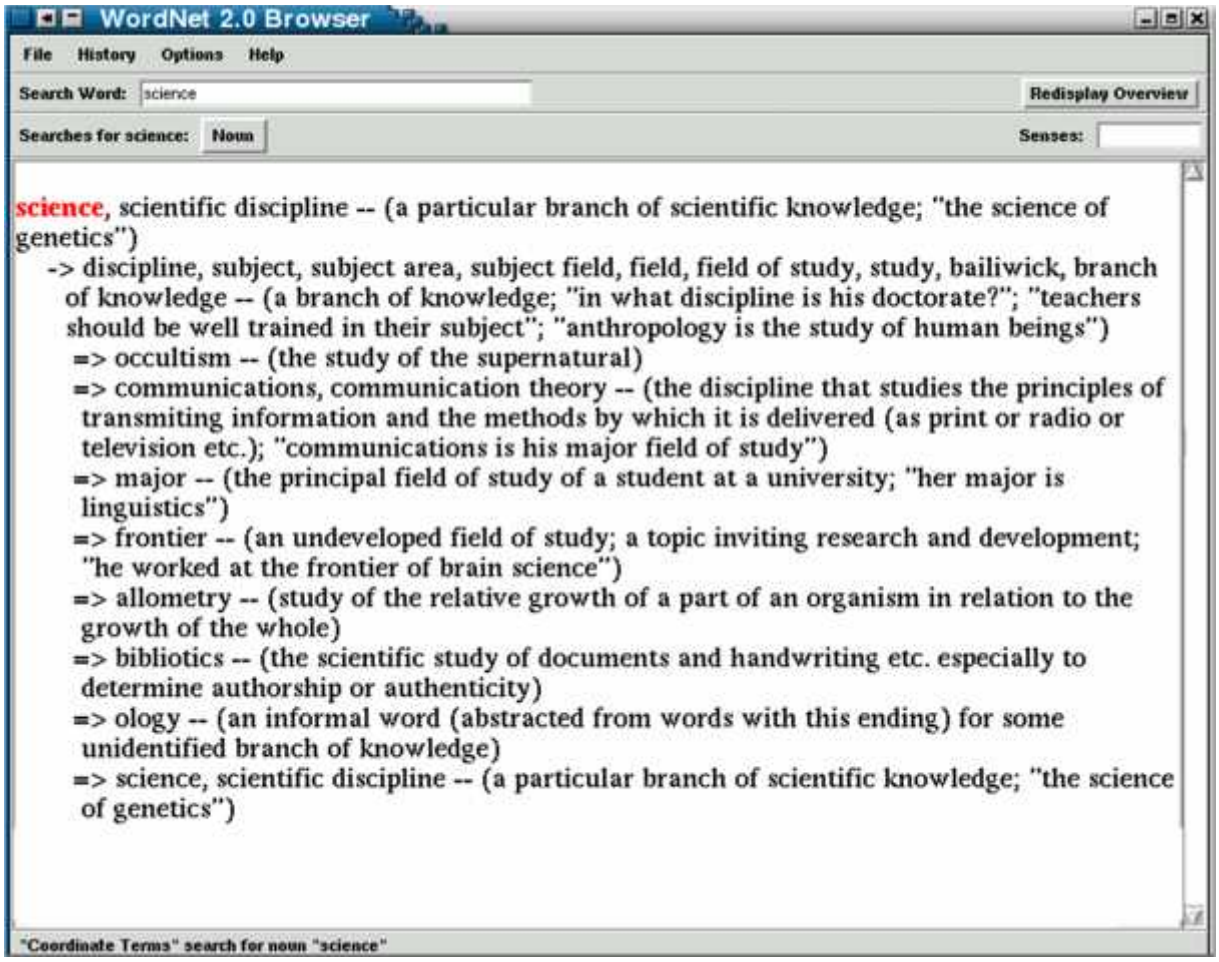


Fig 10 : Exemple de définition de terme en utilisant WordNet

La profondeur d'un terme est le nombre de nœuds le séparant de la racine, dans ce cas la profondeur de 'science' est 8.

Hormis l'efficacité de WordNet, elle représente un problème dans le choix du sens d'un terme.

Par exemple : le terme « java » nous n'avons pas la possibilité de spécifier java le café, la région de l'Island ou bien le langage de programmation.

3.2. Pondération des tags à base du profil utilisateur

Le poids d'un tag est calculé selon l'utilisateur qui l'a mentionné. Ce qui explique qu'un même tag aura deux poids différents dans le cas ou deux utilisateurs le mentionnent, d'une autre part les tags d'un même utilisateur peuvent avoir des poids différents.

Donc le poids de tag est défini en fonction du profil utilisateur représenté par ses dimensions centres d'intérêts et expertise **[Kichou, 10]**.

Le poids d'un tag est calculé selon la formule suivante :

$$W_t^r = \sum_{i=1}^{k} \left[\left(\frac{E \quad (U_i)}{d \quad (I \quad (U_i), P \quad (r))} \right)^{C \quad (U_i, r)} \right]$$

Formule 1 : Pondération de tag a base du profil utilisateur [Kichou, 10]

- k représente le nombre d'utilisateurs ayant annoté la ressource r avec le tag t.

- dist (Interet(Ui), Popularity(r)) désigne la distance entre la ressource composée des tags populaires (vecteur Popularity(r)) et les centres d'intérêts de l'utilisateur (vecteur Interet(Ui)).

Cette distance est calculée comme suit :

$$d \quad (\overrightarrow{I \quad (U_i)}, \overrightarrow{P \quad (r)}) = 1 - \cos(\overrightarrow{I \quad (U_i)}, \overrightarrow{P \quad (r)})$$

Formule 2 : Calcul de la distance entre l'utilisateur et la ressource [Kichou, 10]

Tel que la formule de cosinus est calculée comme suit :

$$\cos\left(\overrightarrow{I \quad (U_i)}, \overrightarrow{P \quad (r)}\right) = \sum_{i=1,K} \frac{W_{i,I}}{\sqrt{\sum_{i=1,K} W_{i,I}^{\ 2}}} \frac{W_{i,P}}{\sqrt{\sum_{i=1,K} W_{i,P}^{\ 2}}}$$

Formule 3 : La mesure cosinus [Gerald, 02]

Avec $W_{i,I}$ est le poids du terme d'indice i dans le vecteur $I \quad (U_I)$ et $W_{i,P}$ est le poids du terme d'indice i dans le vecteur $P \quad (r)$.

- Conf (u, r) représente le degré de confiance d de l'utilisateur u dans son tag, ceci est réalisé avec un rating de 1 à 5 chaque fois qu'un utilisateur taggue une ressource, elle est calculée comme suit :

$$c \quad (u_i, r) = \frac{d}{5}, (d \in \{0,1,2,3,4,5\})$$

Formule 4 : Calcul de la confiance [Kichou, 10]

Par exemple un rating à étoiles : si l'utilisateur ne sélectionne aucune étoiles (pas sûr de son tag) la valeur de la confiance est minimale (0) et le poids calculé devient un simple calcul de popularité, par contre s'il sélectionne 5 étoiles (confiant de ses tags), la valeur de la confiance est maximale (1) et son profil est utilisé dans le calcul du poids du tag.

La valeur de la confiance est identique pour tous les tags assignés pour la ressource par le même utilisateur.

- Exp (u) : l'expertise permet de chercher la part de contribution de la ressource dans l'expertise de l'utilisateur.

Plus cette ressource est proche de l'utilisateur plus la distance *dist(u,r)*est petite et donc le rapport est grand c'est-à-dire qu'un utilisateur tagguant une ressource proche de ses intérêts confirme son expertise , donc lui attribue un poids élevé. Par contre si la ressource diffère de ses centres d'intérêts ne doit pas avoir un grand poids a cause de manque de l'expertise de l'utilisateur dans le domaine.

L'expertise est calculée comme suit :

$$E \quad (U_t) = \frac{\sum p \quad (t_j)}{|T_u|}$$

Formule 5 : Calcul de l'expertise

Donc l'expertise est la moyenne des profondeurs des tags de l'utilisateur où $p \quad (t_j)$ profondeur du tag t_j est le nombre de nœuds le séparant de la racine.

T_u est un ensemble contenant les tags que celui-ci a associé aux ressources, définie comme suit :

$T_u = \{(t_j | U_i, t_j, r) \in Y\}$ avec Y l'ensemble des annotations (actions du tagging).

Le tableau suivant illustre un ensemble de tags et leurs profondeurs :

Tag	Profondeur
Processing	6
Programming	9
Data	5

Tableau 2 : exemple de calcul d'expertise d'un utilisateur

Donc l'expertise d'un utilisateur utilisant ces trois tag (processing, programming, data) en utilisant la formule de calcul de l'expertise (formule 5) est 6,66

3.2.1. Etude des variations de la formule de pondération

Cette étude permet de connaitre les variations du poids exprimé dans la formule 1 en fonctions de ses différents paramètres **[Kichou, 10]**.

- **Variation de la formule de pondération en fonction de l'expertise**

La confiance est fixée à 1 et la distance à 0.1 (c'est-à-dire l'utilisateur est confiant et il est très proche du domaine de la ressource), l'expertise se varie entre 1 et 10, voici la courbe des variations obtenu :

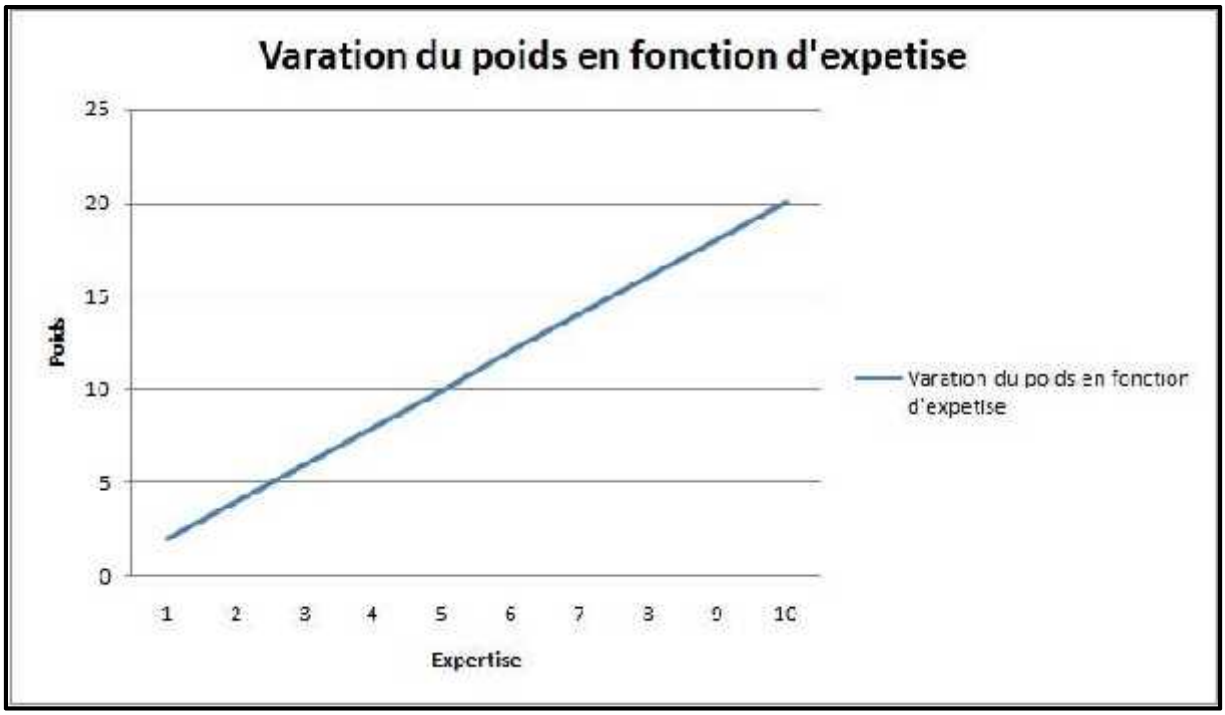


Fig 11 : Courbe des variations du poids en fonction de l'expertise

Le paramètre de l'expertise est très déterminant, car la pondération des tags dépend fortement de ce paramètre.

- **Variation de la formule de pondération en fonction de la distance**

Concernant ce cas, l'expertise a été fixée à 6 et la confiance 0.6 (expertise et confiance moyenne), la distance se varie entre 0.1 et 1, voici la courbe obtenu :

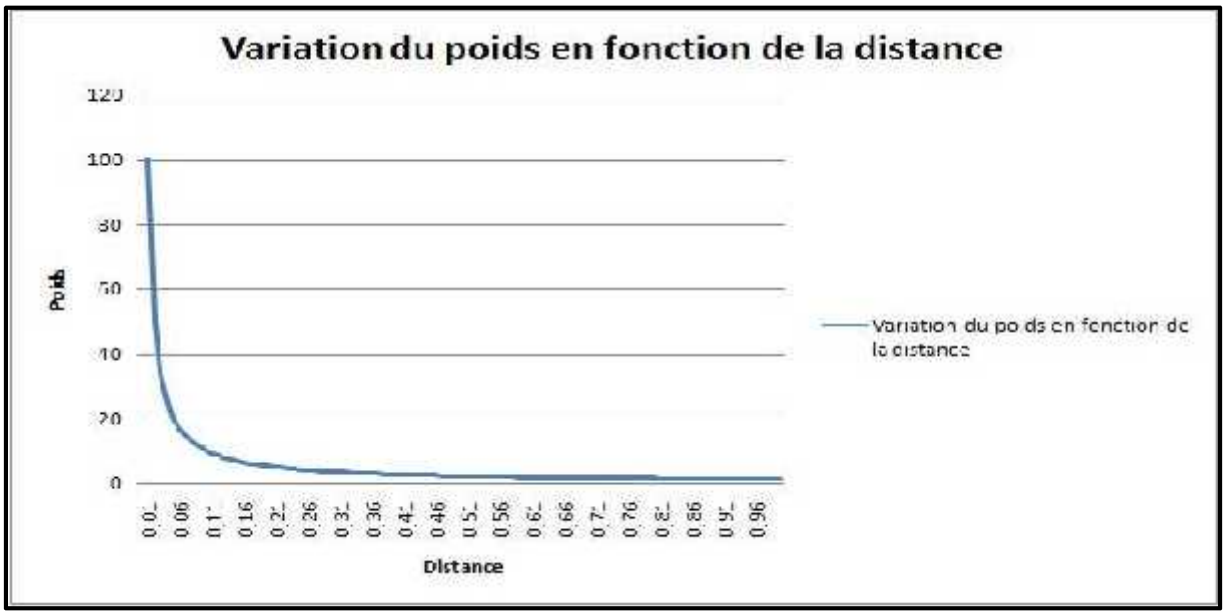


Fig 12 : Courbe des variations du poids en fonction de la distance

La formule de pondération atteint son pic à la valeur minimal de la distance.

- **Variation de la formule de pondération en fonction de la confiance**

Concernant ce cas, le paramètre de l'expertise est fixé à 6 et celui de la distance à 0.5 (utilisateur moyen), par contre la confiance varie entre 0 et 1 (0.2, 0.4, 0.6, 0.8)

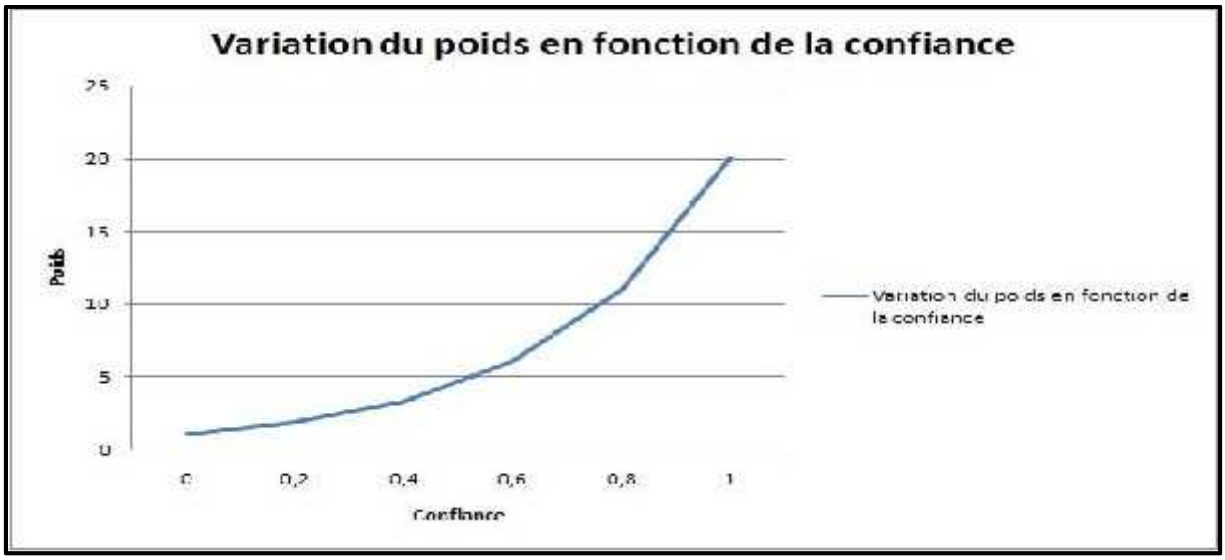


Fig 13 : Courbe des variations des poids en fonctions de la confiance

La formule de la pondération est proportionnelle au paramètre de la confiance mais à un degré moins rapide par rapport au paramètre de l'expertise.

3.3. Classement des tags et construction des descripteurs

Comme c'est déjà mentionné, une ressource donnée a de multiples tags. Une fois que la pondération de ces derniers est faite en utilisant le profil utilisateur, un classement est établi par ordré décroissant de poids des tags.

Par exemple après le calcul des poids des tags d'une ressource donnée nous obtiendrons le tableau suivant :

Tag	Poids
C++	20.12
Programming	19.54
Development	17.65
Languages	10.12
Software	9.87

Tableau 3 : Classement des tags par ordre décroissant

Donc le descripteur de la ressource est un vecteur contenant les tags suivants : *{C++ : 20.12, Programming : 19.54, Developement : 17.65}*

Dans cet exemple il n'y a que trois tags sélectionnés vu que le tableau ne contient que cinq tags en tout, mais en réalité il y a des multiples tags utilisés et le choix des descripteurs dépend du nombre total des tags concernant une ressource.

4. Les points forts de l'approche

L'approche du filtrage à base du profil utilisateur permet de calculer les poids des tags en incluant le profil utilisateur c'est-à-dire s'intéresse aux centres d'intérêts et l'expertise de l'utilisateur, contrairement aux sites actuels qui se basent seulement sur le principe de la popularité des tags.

Donc cette approche est plus avantageuse dans la création des descripteurs des ressources, car le calcul du poids d'un tag en prenant en considération le profil de l'utilisateur permet de pondérer les tags qui décrivent la ressource d'une manière plus précise, ce qui est expliqué par l'établissement d'un nouveau classement des tags plus corrects.

5. Les points faibles de l'approche

Malgré que l'approche de filtrage basée sur le profil d'utilisateur représente des avantages par rapport à d'autres approches en donnant un descripteur de ressource plus déterminé , elle a certaines limites, parmi ses limites nous citons :

- **Problème d'ambigüité :** c'est-à-dire que l'approche du filtrage ne fait pas la différence entre deux mots qui s'écrivent de la même manière mais qui ont un sens différent exemple : en calculant le poids d'un tag d'une ressource donnée en utilisant l'approche du filtrage et le mot 'Orange' soit désigner comme un descripteur, nous nous ne pourrons pas savoir s'il s'agit d'un fruit, couleur où l'opérateur mobile.
- **Problème de variation d'écriture :** l'approche ne fait pas aussi la différence entre divers formes d'un même mot c'est-à-dire si un utilisateur associe le tag 'web2.0' à une ressource donnée et un autre utilise 'web2-0' pour la même ressource, l'approche du filtrage ne considère pas ces deux tags comme identique et calcul le poids de chacun d'eux.
- **Le calcul d'expertise :** l'approche du filtrage se base sur l'ontologie *Wordnet*, alors qu'une ontologie de domaine peut affecter des résultats plus précis.
- **La confiance :** l'utilisateur représente le degré de confiance par rapport à la ressource, c'est-à-dire qu'il évalue chaque ressource avec un rating de 1 à 5, mais un utilisateur confiant par rapport à la ressource, peut l'être pour un tag mais pas pour les autres.
- **Utilisation limitée :** l'approche de filtrage à base du profil utilisateur est validée seulement sur une collection moyenne comportant : 149 URLs, 215 tags, et 565 actions de Tagging réalisées par 6 utilisateurs.

6. Les solutions existantes pour remédier au problème d'ambigüité

Malgré le succès et la force des folksonomies, celles-ci présentent des problèmes dus aux vocabulaires non contrôlés adoptés par les systèmes du tagging collaboratif notamment le problème d'ambigüité, pour cela un certain nombre de travaux se sont penchés sur ce problème de l'annotation des documents et du partage des connaissances.

6.1. Les approches basées sur les ontologies

Ces approches sont soit en guidant le tagging à l'aide d'ontologie, où bien en construisant une ontologie de folksonomie.

6.1.1. Guider le tagging à l'aide d'une ontologie

De nombreuses solutions visent à intégrer des systèmes à base d'ontologies ont récemment vu le jour.

[Passant, 07] l'ontologie subordonne l'annotation en aiguillant le choix du tag : il s'agit en fait pour l'utilisateur de réutiliser un tag existant, ou de proposer un nouveau tag pour une ressource qui existe déjà. Cependant l'utilisateur reste libre de soumettre une ressource et son tag s'il n'existe pas encore. L'intérêt vient du fait que l'ambigüité des tags par rapport aux ressources qu'ils sont censés désigner est automatiquement levée.

6.1.2. Construire une ontologie de folksonomie

[Gruber, 07] propose de construire une ontologie de folksonomie. La "TagOntology" est un projet de construction d'une ontologie commune dédiée à la formalisation et la conceptualisation de l'acte d'annoter une ressource par un terme (tagging). Ce modèle met en œuvre quatre entités :

- l'objet tagué (la ressource).
- le tag (c'est `a dire le mot-clé utilisé).
- l'utilisateur taguant.
- le domaine au sein duquel le tagging est affecté.

[Gruber, 07] va plus loin dans la réification des tags et considère chaque tag comme un objet à part entière. Pour contrer les ambiguïtés des tags ou les usages abusifs (spam), **[Gruber, 07]** proposait déjà de « tagger les tags » comme ça il serait possible d'indiquer que tel tag est synonyme de tel autre tag, ou encore que tel tag est adéquat ou non pour tel objet.

6.1.3. Utiliser le tagging pour consolider les tags (tags4tags)

L'idée de base est d'étendre le rang de l'objet principal du tagging qui est 'la ressource' à une union des ressources et des paires de tags (T_1, T_2) , désormais l'utilisateur peut non seulement tagguer des ressources mais aussi des tags avec d'autres tags recommandés par le système ou de son choix tel que 'est un', 'signifie', comme ça permet de tagguer les relations entre les tags par exemple T_1 (file) et T_2 (fichier) , un utilisateur peut tagguer la relation (T_1, T_2) en associant le tag 'Anglais-Français' **[Garcia-castro et al, 09]**.

6.2. Exemples d'ontologies d'informatique pour le tagging

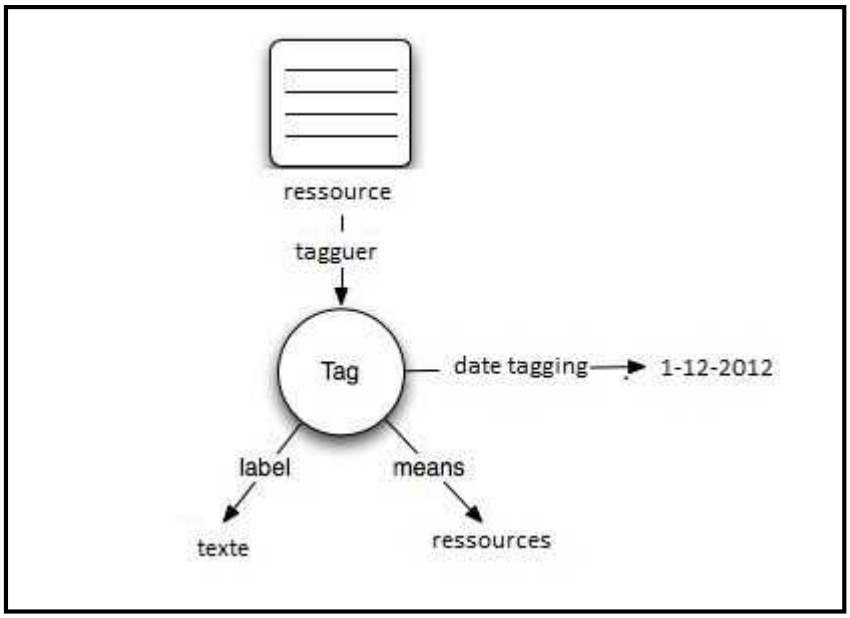
Plusieurs ontologies ont été conçues pour guider le tagging collaboratif, parmi ces ontologies nous citons :

6.2.1. Common tag

Le format "Common Tag" a été développé par un groupe d'agences Web (des entreprises qui font du Web) qui aide les éditeurs à coups de tagging sémantique pour permettre de rendre leur contenu encore plus riche. Cela devrait donc contribuer à rendre ces ressources web plus facile à trouver et à réutiliser.

La structure de Common Tag est relativement simple. Une ressource, accessible via une URL (adresse internet unique) peut être décrite avec des tags. Chaque tag pointe vers une autre ressource qui identifie le concept décrit. Un tag peut éventuellement contenir d'autres informations comme la date du tagging et champs texte. **[CommonTag]**.

La structure de common tag est représentée dans la figure suivante :



.

Fig 14 : Structure de common tag [Commontag]

6.2.2. MOAT

[Passant et Laublet, 08] dans son article les ontologies du web 2.0 parle de MOAT (Meaning Of A Tag) qui est un système permettant de préciser la signification des tags utilisés pour catégoriser des contenus d'une ressource.

Le but de ce système est de faire :

- Une solution pour résoudre les problèmes de « free-tagging » (ambiguïté, hétérogénéité, le manque d'organisation).
- Une ontologie pour représenter sens global et local des étiquettes d'une manière compréhensible par une machine.
- Un cadre d'assigner et de partager des significations aux mots d'une manière ouverte et collaborative.
- Un processus de création de données liées de simple actions de marquage « tagging».

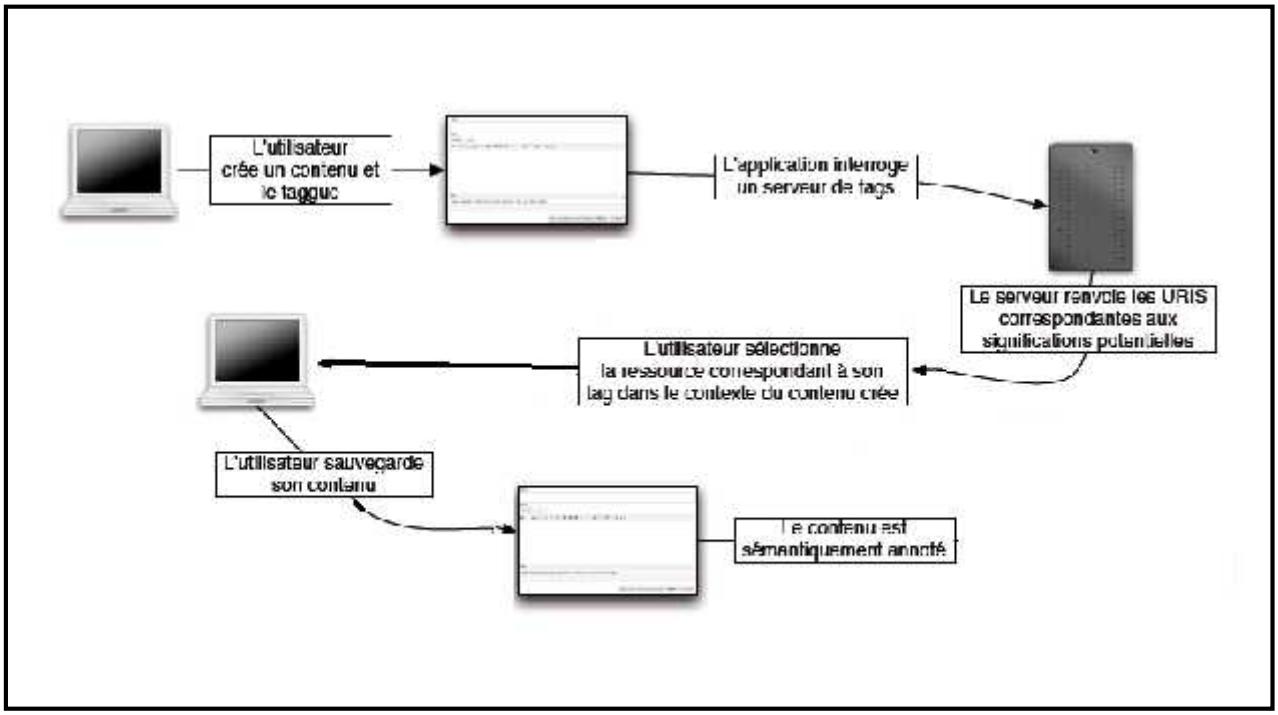


Fig 15 : Processus de communication entre un client et un serveur MOAT [Passant et Laublet, 08]

7. Les solutions existantes pour remédier au problème de variation d'écriture

Notamment le problème de variation d'écriture est une autre limite causée par la liberté des choix des tags qu'offre le système du tagging aux utilisateurs, cependant plusieurs solutions ont été proposées pour remédier à cette limite.

7.1. Clustering

La méthode de **[Specia et Motta, 07]** consiste à regrouper les tags fortement liés entre eux dans un même cluster.

La première étape de cette méthode s'agit de construire une matrice de cooccurrence qui rassemble les tags de la même ressource afin d'élaborer des clusters des tags. La deuxième étape consiste à expliciter les liens qui existent entre les tags d'un même cluster.

7.2. Détection des variations d'écriture

[Levenshtein, 96] a utiliser une méthode pour mesurer la distance entre deux tags afin de détecter les variations d'écriture des tags supposés équivalent.

Le tableau suivant (Tableau 4) montre les valeurs de la distance mesurée pour certains couples de tags.

Tag 1	Tag 2	Distance de Levenshtein
Email	E-mail	1
Informatique	Information	3

Tableau 4 : Distance de Levenshtein pour certains couples de tags.

Donc la distance de levenshtein est le nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaine de caractère à une autre.

8. Conclusion

Le tagging collaboratif permet aux utilisateurs d'associer librement des mots clé à des ressources partagées ce qui implique qu'un tag ne peut pas forcement représenter le contenu. En sachant que les tags sont classés par ordre de popularité, cette liberté de choix des tags causera de nombreux problèmes, particulièrement en ce qui concerne l'attribution des tags représentatifs.

Donc pour pallier à ces problèmes une approche de filtrage a été proposée et celle-ci est basée sur le profil utilisateur et qui a pour objectif de créer un descripteur de ressource assez représentatif.

Dans ce chapitre nous avons expliqué par donner le principe général de cette approche, ensuite nous avons présenté l'approche ainsi que ses étapes, la modélisation du profil, la pondération des tags et la création de descripteur de ressource, par la suite nous avons assigné

ses avantages et inconvénients, enfin nous avons mentionné quelques solutions qui ont été proposées pour remédier à certaines limites.

Malgré l’efficacité de l’approche du filtrage des tags à base du profil utilisateur dans la création des descripteurs des ressources, elle représente toujours certaines limites, que nous devons corriger en choisissant les meilleures solutions à réaliser, celles-ci seront abordées dans le prochain chapitre.

Deuxième Partie
Contribution

CHAPITRE III : Conception du système

1. Introduction

Suite à l'évolution du web 2.0 de nouvelles fonctionnalités ont vu le jour, parmi celle-ci le tagging collaboratif qui laisse la liberté totale aux utilisateurs d'associer des mots clés à des ressources données, cette liberté provoque certains inconvénients notamment aux choix des descripteurs de ressources, car la plupart des sites considèrent les tags les plus populaires étant un descripteur pour la ressource, alors qu'il ce peut que ce tag ne décris et ne représente même pas bien la ressource concernée.

Pour résoudre ce problème une approche de filtrage de tags à base du profil utilisateur a été proposée dans **[Kichou, 11]** permettant de calculer le poids des tags selon les centres d'intérêts des utilisateurs ainsi que leurs expertises, ce qui permet d'avoir des descripteurs beaucoup plus conforme à la ressource mais cette approche n'a pas traité les problèmes d'ambigüité et variations d'écritures des tags ainsi que d'autres limites que nous avons citées dans le chapitre précédant.

Dans ce chapitre nous allons présenter les solutions choisies afin de pallier aux limites de l'approche. D'abords nous parlerons des motivations qui nous ont poussées à choisir cette problématique qui permet d'apporter des améliorations aux travaux existant, ensuite nous allons citer les solutions à réaliser, ainsi que la conception du système en intégrant ces solutions et enfin nous allons donner quelques exemples illustrant le travail proposé.

2. Motivations

Notre travail rentre dans le domaine du tagging collaboratif, ce moyen est devenu de plus en plus populaire sur le web. Chaque jour plusieurs utilisateurs s'inscrivent, en ayant la possibilité de tagguer de multitude contenus partagés. La majorité des sites de tagging considèrent les tags populaires autant que descripteurs de ressources, sachant qu'ils ne les représentent pas nécessairement.

[Kichou, 10] a proposée de calculer le poids de chaque tag en se basant sur le profil utilisateur, ce qui permettra d'avoir des tags représentant au mieux la ressource concernée. Mais n'empêche que cela n'écarte pas certains problèmes comme :

- **L'ambigüité** : deux mots identique mais de différent sens.
- **La variation d'écriture** : un mot peut être écrit sous diverses formes.
- **La confiance :** l'attribution de la confiance par rapport à la ressource.

Ainsi nous avons le problème de l'ontologie qui est trop vaste (Wordnet), nous proposons qu'elle soit spécifique à un domaine pour avoir des résultats plus exacts.

3. Choix des solutions à réaliser

Dans le chapitre précédant nous avons abordé les points faible de l'approche de filtrage à base du profil utilisateur, ce qui nous à conduit à proposer les solutions suivantes pour pallier à ces limites :

3 .1. Solution proposée pour la variation d'écriture

Les utilisateurs peuvent attribuer un même tag à une ressource sous différentes formes (exemple : email, e-mail, e.mail…), l'approche de filtrage à base du profil utilisateur considère les tags écrits sous diverses formes comme étant des tags différents et calcul le poids de chacun d'eux. Cependant nous avons proposé la solution suivante afin d'identifier les tags identiques, cette solution est constituée des étapes suivantes :

- Stemmatisation des termes de l'ontologie

- Calcul de la distance entre deux termes (damerau-levenshtein)

- La suggestion des termes

1. Stemmatisation (racinisation) des termes de l'ontologie

La stemmatisation est une technique qui vise à transformer les mots en leur radical, elle cherche selon le mot et la langue et définit le radical le plus probable pour ce mot, elle fonctionne uniquement avec une base de connaissance des règles syntaxiques et grammaticales de la langue.

Le stemme (racine) c'est la partie restante d'un mot une fois que son suffixe et préfixe sont supprimés, il peut ne pas correspondre à un mot réel **[Lovins, 68]**.

Par exemple :

Mot	Stemme
Biological	Biolog
Information	Inform

Tableau 5 : Exemple de stemmatisation des termes

Comme nous l'observons dans le tableau (tableau 5) le mot « biolog » n'est pas un mot réel.

Donc la première étape consiste à stemmatiser tous les termes de la base de données, ainsi que leurs synonymes, pour cela nous allons utiliser l'algorithme de *porter stemmer* pour sa simplicité et sa rapidité.

- Quelques algorithmes de stemmatisation

Voici quelques algorithmes permettant la stemmatisation des mots:

- **Porter stemmer (anglais)**

Cet algorithme est l'un des plus populaires algorithmes de stemmatisation développé par **[Porter, 80]**, sa principale utilisation est dans le cadre d'un processus de normalisation des termes qui est effectué généralement lors de la mise en place d'un systéme de récupération d'informations. Plus précisément l'algorithme comporte cinq étapes, chaque étape définit un ensemble de règles. Pour stemmatiser un mot, les règles sont testées séquentiellement, comme le montre la figure suivant :

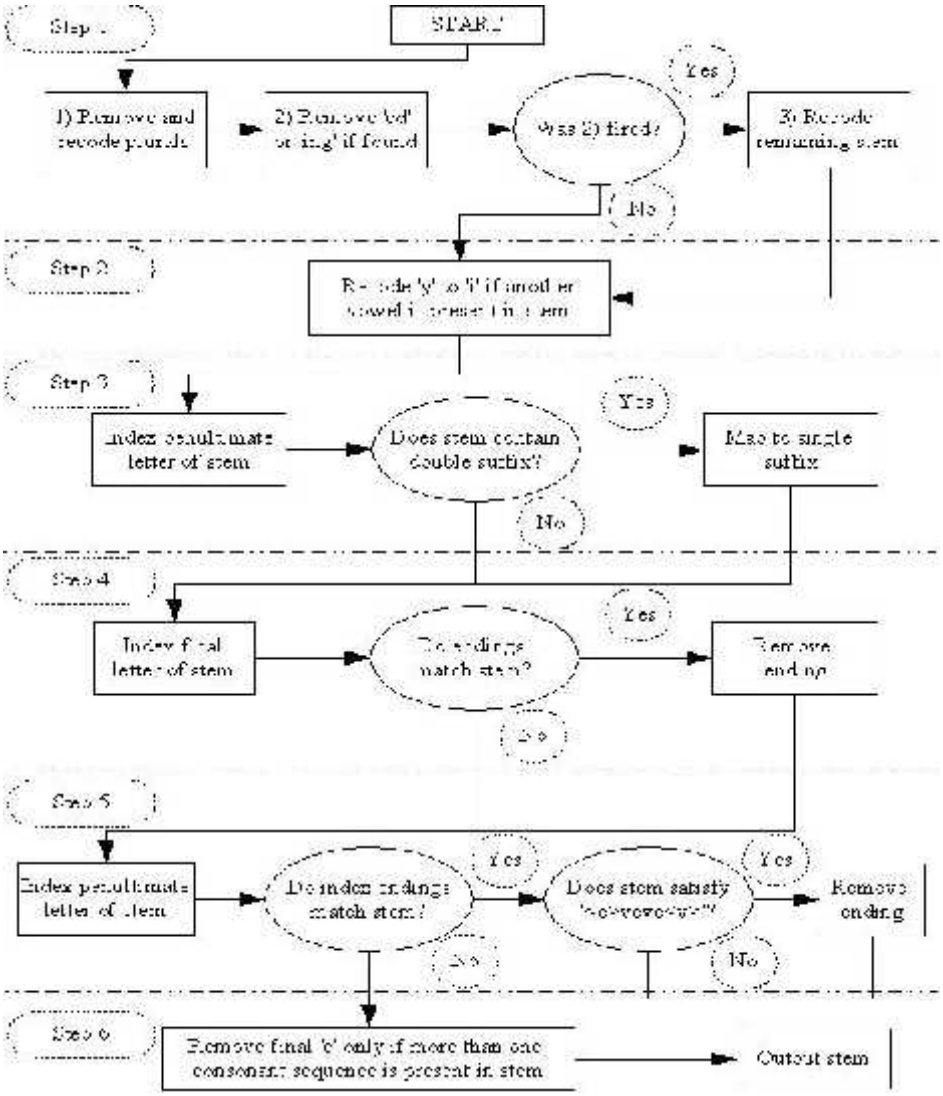


Fig 16 : les étapes de l'algorithme de porter stemmer [Porter, 80].

Les suffixes d'un mot sont retirés par étapes, la première étape de l'algorithme consiste à éliminer les pluriels et les terminaisons des verbes conjugués dans différents temps (présent, participe passé et participe présent) par exemple :

- Biologicals ⟶ Biological.
- Implemented ⟶ Implement.
- Programing ⟶ Program.

Dans la deuxième étape si la dernière lettre d'un mot est « y », alors elle sera transformée en un « i », seulement si une autre voyelle est présente dans le mot, par exemple :

- Biology ⟶ Biologi.

Quant à la troisième étape gère les mots contenant un doubles suffixes, en supprimant un et avoir un mot avec un seul suffixe, l'étape suivante traite les doubles suffixes non prise en charge par l'étape précédente, ensuite dans la cinquième étape les simples suffixes sont supprimés, voir le tableau suivant :

Etape	Mot	Déroulement
3	Generalization	Generalize
4	Generalize	General
5	General	Gener

Tableau 6 : exemple illustrant les étapes « 3, 4, 5 » de l'algorithme de porter

Enfin la sixième étape et la dernière traite les mots contenant plus d'une consonne et qui se terminent par un « e », par exemple : Compute ⟶ Comput.

Vu la fiabilité de cet algorithme nous avons décidé de l'employer afin de stemmatiser les termes de l'ontologie.

- **Carry stemmer (français)**

C'est un algorithme permettant la désuffixation dans la langue française toute en suivant le modèle de **[Porter, 80]**.

Cet algorithme se déroule selon diverses étapes en passant successivement les mots à analyser, il est constitué d'un analyseur morphologique qui consiste à reconnaitre les suffixes d'une liste des mots, soit il les supprime soit il les transforme **[Paternostre et al, 02]**.

- **Paice / Husk stemmer**

Un algorithme permettant de faire la stemmatisation, se base sur une seule table de règles pour extraire la racine, ces règles sont stockées en dehors du code source, ce qui permet d'utiliser cet algorithme dans différentes langues juste en changeant certaines règles. Ainsi l'algorithme est facilement utilisé pour la gestion de plusieurs langues **[Paice, 90]**.

2. Calcul de la distance entre deux termes (damerau-levenshtein)

Il existe plusieurs méthodes afin de calculer la distance entre deux chaines de caractères parmi celles-ci il y a la distance de *damerau-levenshtein* pour qui nous avons apporté

quelques modifications dans l'intention de l'utiliser dans la réalisation de la solution que nous avons proposée.

La distance *damerau-levenshtein* est une distance entre deux chaines de caractères, son principe est de calculer le nombre minimal d'opérations nécessaire pour transformer une chaine de caractère vers une autre, ces opérations sont : l'insertion, suppression, substitution d'un caractère et la transposition de deux caractères **[Brill et Moore, 00]**.

Voici la distance entre certains couples de termes illustrée dans le tableau suivant :

Terme 1	Terme 2	Distance damerau-levenshtein
Medical	Medicals	1
Systems	System	1
Biological	Biology	4

Tableau 7 : exemple distance de damereau-levenshtein entre certains couples de termes

Comme nous l'observons dans le tableau ci-dessus, le nombre d'opération effectuée afin de transformer le mot « Medical » vers « Medicals » est égal à 1, en insérant une lettre.

Donc dés qu'un utilisateur attribue un tag à une ressource, nous générons tous les mots qui ont une distance égale à 1 si l'une des opérations de substitution, insertion, transposition a été effectués une et une seule fois de manière exclusive, quant à la suppression nous permettons son exécution deux fois par rapport au tag attribué.

3. La suggestion des termes

Une fois la distance calculée et les mots généré, nous allons stemmatiser tous ces mots et les comparer avec les stemmes des termes de l'ontologie, si deux stemmes sont identiques nous suggérons à l'utilisateur la liste des termes (et leurs synonymes) équivalents à son tag, ce qui écartera le problème de la variation d'écriture des tags et aidera à minimiser l'ambigüité.

3.2. Solution proposée pour l'ambigüité

L'ambigüité est le fait qu'un mot ait plusieurs sens (exemple : le mot « avocat » signifie soit un fruit soit une profession), cette ambigüité cause plusieurs problèmes dans les systèmes du tagging, notamment dans les choix des descripteurs des ressources et aussi dans le retour

des résultats de recherche, à cet effet nous avons proposé de guider le tagging par une ontologie de domaine afin d'éliminer le problème d'ambigüité.

Le déroulement des différentes étapes afin d'éliminer la variation d'écriture des tags contribue à mettre fin au problème d'ambigüité

D'une autre façon, si un utilisateur attribue un tag ambigu, le fait de lui suggéré les synonymes du tag, l'utilisateur pourra ainsi préciser le sens exacte de ce dernier, ce qui guidera à enlever l'ambigüité.

- **Ontologie du domaine Biomédical**

L'ontologie que nous allons nous servir est de domaine biomédicale (L'ontologie est présentée dans l'annexe B), c'est l'une des bases de données la plus complète, elle offre un environnement unique pour la recherche médicale, elle est constituée de plusieurs sous ontologie médicale, parmi celles-ci :

- **Newt uniprot taxonomy :** est une base de données centrale de séquences de protéines avec précision et cohérence, elle contient 1025199 termes de ce domaine.

- **Gene ontologie :** est une ontologie génétique qui a développée trois vocabulaires contrôlés et structurés : les composants cellulaires, les processus biologiques et les fonctions moléculaires, elle est constituée de 36455 termes.

La figure suivante illustre la profondeur d'un terme dans l'ontologie biomédicale :

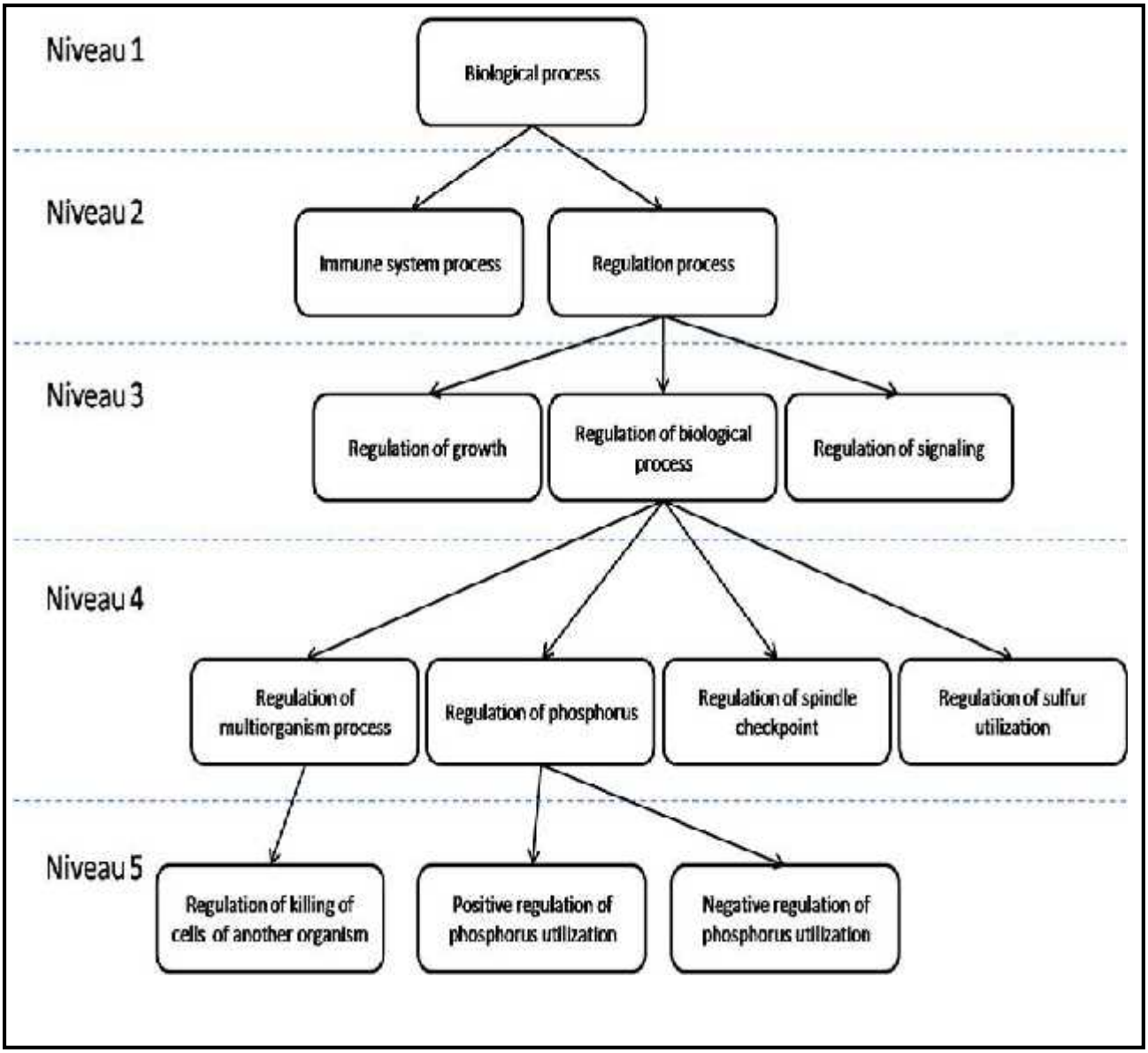


Fig 17: Exemple de profondeur d'un terme avec l'ontologie biomédicale

3.3. La solution proposée pour la confiance

Dans l'approche de filtrage de tags à base du profil utilisateur, la confiance est accordée par rapport à la ressource, or un utilisateur confiant vis-à-vis une ressource, ne garanti pas qu'il aille affecter des tags significatifs à cette dernière.

Alors nous avons proposée aux utilisateurs d'approprier un degré de confiance pour chaque tag attribué, ce qui nous permettra d'avoir des descripteurs de ressources beaucoup plus spécifiques.

4. Conception des solutions

Le système que nous proposons concerne l'action du tagging collaboratif et plus précisément c'est une extension de l'approche du filtrage à base du profil utilisateur.

Dans ce cas nous allons présenter l'architecture du système ainsi que la base de données, pour permettre de clarifier les solutions que nous avons proposé.

4.1. L'architecture du système

Nous illustrons notre système par un schéma global (fig18) permettant de détailler les améliorations apportées à l'approche de filtrage des tags à base du profil utilisateur.

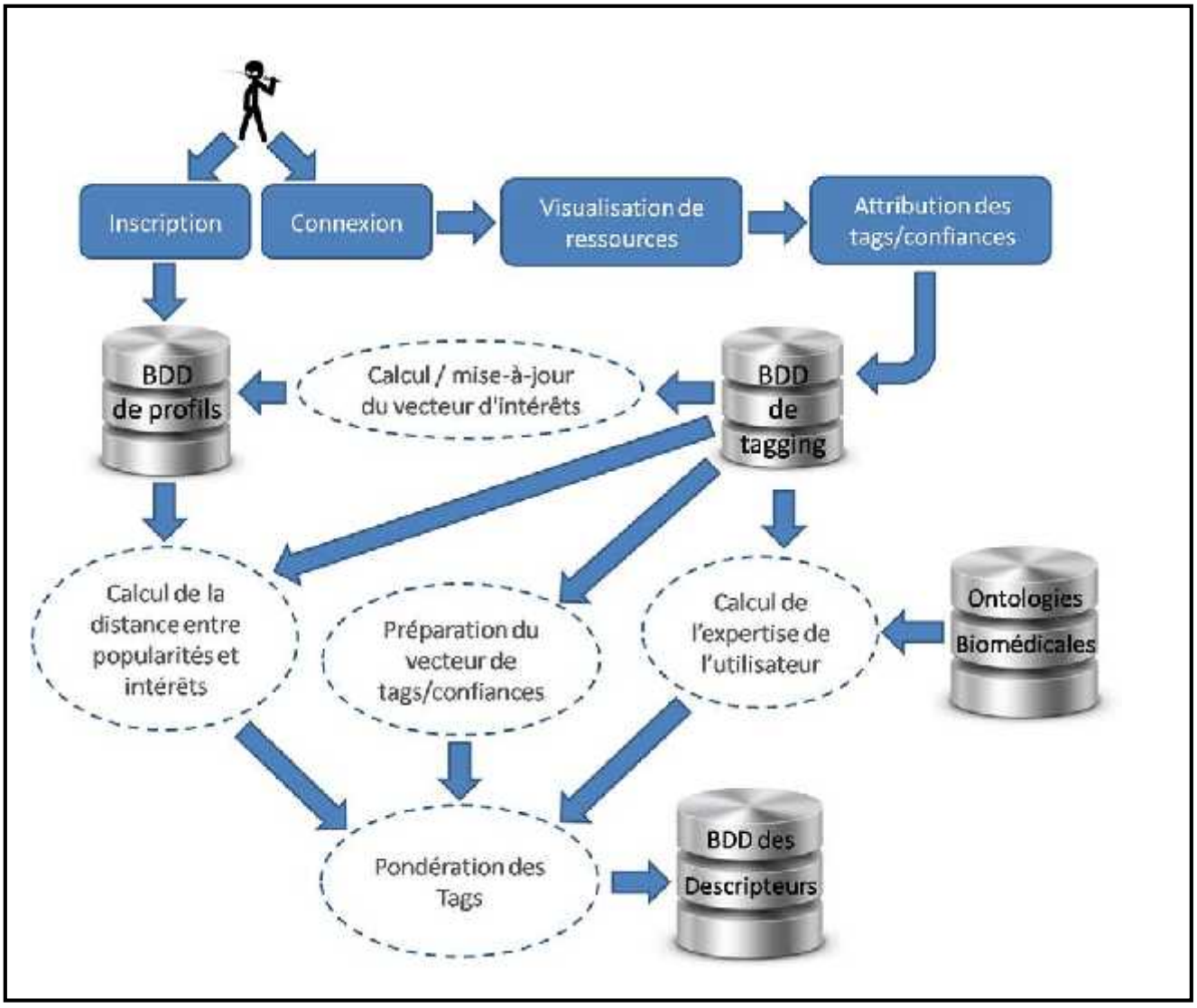


Fig 18 : Le schéma global

Le schéma ci-dessus montre le déroulement des différentes étapes afin de pondérer les tags d'une ressource, en premier lieu il y a le calcul et la mise à jour du vecteur d'intérêts d'un utilisateur, ensuite le calcul de l'expertise, la préparation du vecteur de tags avec la confiance de chaque tag attribuée par l'utilisateur, ainsi que le calcul de la distance entre le vecteur de popularité et le vecteur d'intérêts, une fois ces étapes effectuées, un classement décroissant de tags par leurs poids est affiché, les meilleurs tags sont sélectionnés pour former les descripteurs.

Afin de mieux préciser la conception de notre système nous avons proposé un diagramme de cas d'utilisation dans la figure suivante :

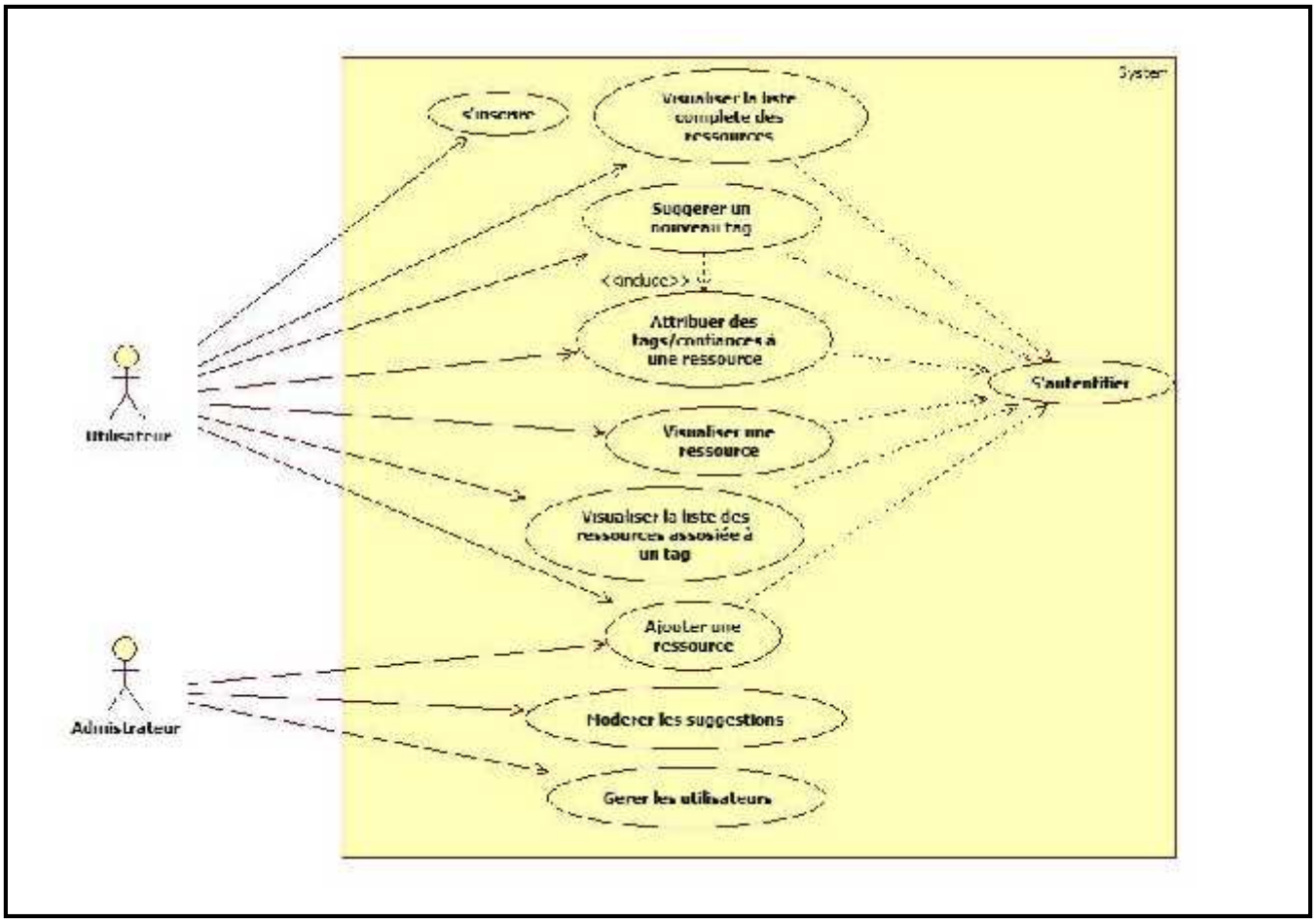


Fig 19 : Diagramme de cas d'utilisation

Les différentes fonctionnalités d'un utilisateur sont :

- **Inscription :** un utilisateur peut s'inscrire dans notre système en introduisant ses informations personnelles (nom, prénom, mot de passe, date de naissance, e-mail …).
- **Ajout d'une ressource :** chaque utilisateur inscrit a le droit d'ajouter une ressource quelconque.

- **Attribution des tags :** l'annotation des ressources se fait librement par n'importe quel utilisateur, mais il à la possibilité de choisir parmi les tags que notre système lui suggère, pour permettre d'avoir des tags exacte, ainsi l'utilisateur pourra donner le degré de confiance par rapport à son tag.
- **Visualisation d'une ressource :** les ressources partagées sur notre système sont visualisées par la totalité des utilisateurs inscrits.
- **Visualisation de liste de ressource associée à un tag :** le système permet à l'utilisateur de visualiser les ressources concernant un tag précis, par exemple : les ressources qui ont un rapport avec le tag « java ».
- **Visualisation de la liste complète des ressources :** le système permet aux utilisateurs d'afficher la liste complète des ressources partagées.
- **Suggestion des tags :** un utilisateur à la possibilité de suggérer un tag qui n'existe pas dans l'ontologie.

Les principales fonctionnalités de l'administrateur :

- **la modération des suggestions de tag :** l'administrateur vérifie tout les tags suggérés par les utilisateurs, il peut soit les validés en les ajoutant à la base de données ou bien les rejetés.
- **Ajouter une ressource :** comme tout utilisateur, l'administrateur à la possibilité d'ajouter de nouvelles ressource.
- **Gérer les utilisateurs :** l'administrateur peut visualiser la liste des utilisateurs inscrits, en ayant accès à leurs informations personnelles.

4.2. L'architecture de la base de données

Notre diagramme de classe comporte les tables suivantes (fig 20) :

- **Table Ressource :** contient l'ensemble des ressources partagées par les utilisateurs définies par un titre, un résumé et un url.
- **Table Utilisateur :** contient l'ensemble des utilisateurs inscrits dans le système en introduisant leurs noms, prénoms, e-mails, mots de passe, sexe et date de naissance.
- **Table Tagging :** représente l'action du tagging, contient le degré de confiance de chaque tag attribué par l'utilisateur par rapport à une ressource.

- **Table Term :** représente les termes de l'ontologie avec leurs définitions et leurs stemmes.
- **Table Synonyme :** contient l'ensemble des synonymes des termes et leurs stemmes.
- **Table PoidsPoularité :** inclut le poids et la popularité de chaque terme attribué à une ressource.
- **Table suggestion :** elle comporte les nouveaux tags avec leurs confiances suggérés par les utilisateurs par rapport à une ressource.

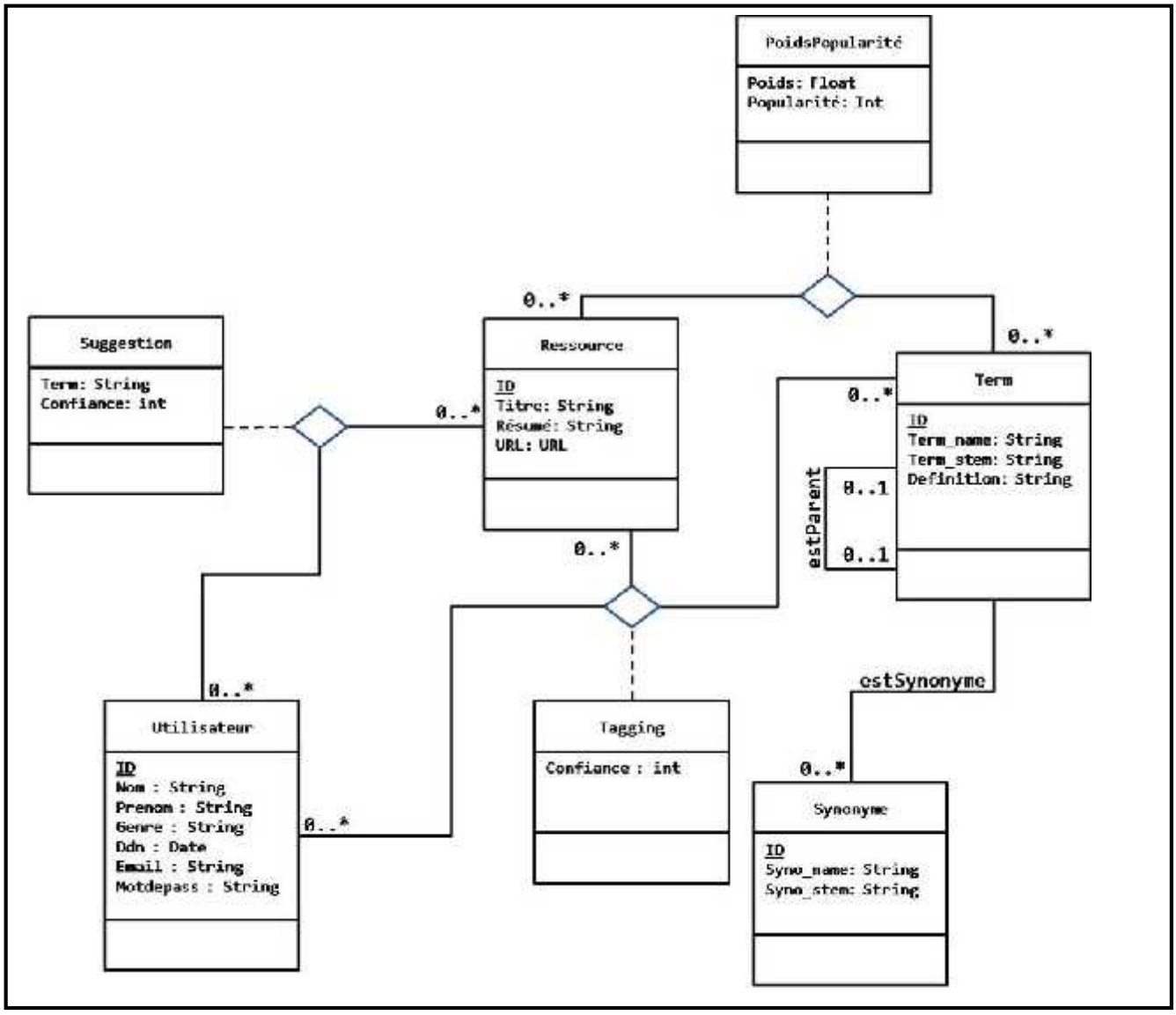


Fig 20 : diagramme de classes

5. Exemple illustrant l'apport de la version étendue

Après avoir détaillé les solutions que nous avons choisies afin de régler les difficultés empêchant le bon fonctionnement de l'approche de filtrage des tags à base du profil utilisateur dans le choix des descripteurs des ressources, nous allons montrer dans l'exemple suivant les tags des dix utilisateurs associés à la ressource « *insulin resistance resistance in pulmonary arterial hypertension* »,

Sachant que le vecteur de ses tags les plus populaires est : [HIV : 8, heart failure : 8, insulin resistance : 6, hypertension : 5, death : 3].

La figure suivante montre les centres d'intérêts des utilisateurs calculés par l'approche hybride (naïve/cooccurrence) :

U1:[	symptom(15)	death(9)	insulin resistance(20)	HIV(3)	protein(9)	]
U2:[	HIV(15)	obesity(13)	death(5)	symptom(10)	analysis(8)	]
U3:[	symptom(22)	conical(5)	sick(12)	artery(7)	cardiomyopathy(2)	]
U4:[	cardiomyopathy(10)	obesity(2)	hypertension(6)	analysis(11)	artery(4)	]
U5:[	HIV(14)	heart failure(13)	cardiomyopathy(16)	pathogens(10)	diabetes(8)	]
U6:[	HIV(12)	hypertension(10)	sick(14)	peroxisome(6)	conical(6)	]
U7:[	diabetes(6)	pathogenesis(8)	cardiomyopathy(5)	obesity(2)	artery(7)	]
U8:[	HIV(4)	heart failure(7)	peroxisome(10)	sensitization(2)	protein(4)	]
U9:[	transcription(9)	secretion(11)	phosphorylation(14)	behavior(2)	sensitization(6)	]
U10:[	fever(10)	sleep(18)	transcription(14)	Insulin resistance(10)	apoptosis(6)	]

Fig 21 : Centres d'intérêts des utilisateurs

La distance entre le vecteur d'intérêt des dix utilisateurs et le vecteur de popularité ainsi que l'expertise sont représentés dans le tableau suivant :

Utilisateurs	Distance	Expertise
U1	0,49	2,25
U2	0,46	1,88
U3	0,86	3
U4	0,72	2
U5	0,33	2
U6	0,54	4
U7	0,89	1,75
U8	0,54	3,22
U9	1	3,78
U10	0,84	2,7

Tableau 8: Distance, expertise des dix utilisateurs

Le tableau ci-dessous résume les tags attribués par les utilisateurs avec leur degré de confiance, ainsi que leurs popularités et leurs poids.

Tag	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	Popularité	Poids
HIV	0,2	0,4	0,6	0,4	0,2	0	0,2			0,2	8	11,54
Heart failure	0,6	0,2	1	0,4	0,4		0,2		0,4	0,2	8	14,92
Insulin resistance	0,6	0,6				1	0,8	0,4	0,6		6	18,12
Hypertension	0,8	1	0,6			0,6			1		5	16,56
Death					0,2	0,2				0,2	3	4,18
cardiomyopathy	0,2							0,2	0,4		3	4,47
Symptom								0,2		0,6	2	3,15
Obesity								0,4		0,8	2	4,57

Tableau 9: Liste des tags associé à une ressource

Enfin nous allons faire une comparaison entre les tags populaires et les tags pondérés en calculant leurs poids en intégrant le profil utilisateur (tableau 10).

La popularité des tags	Le poids des tags
HIV (8)	Insulin resistance (18,12)
Heart failure (8)	Hypertension (16,56)
insulin resistance (6)	Heart failure (14,92)
Hypertension (5)	HIV (11,54)
Cardiomyopathy (3)	Obesity (4,57)

Tableau 10: Comparaison entre la popularité et le poids des tags.

Nous remarquons que la liste des tags populaire est différente de celle des tags pondérés en prenant en considération le profil utilisateur, et cela parce que l'expertise de l'utilisateur ainsi que ses centres d'intérêt jouent un rôle important dans le choix des descripteurs des ressources.

Donc le nouveau vecteur des descripteurs de la ressource est : [insulin resistance (15,08), hypertension (14,64), heart failure (13,1), HIV (10,65), cardiomyopathy(4,45)], qui est beaucoup plus significatif que le vecteur des tags populaires, sachant que le tag « insulin resistance » est plus spécifique à la ressource que le tag « HIV », également pour « hypertension ».

D'un autre coté notre calcul donne très rarement des tags de même poids, par exemple : les tags « death » et « cardiopathymy » qui ont une popularité similaire, leurs poids est différent (tableau 10).

Dans cet exemple nous avons pris que les cinq premiers tags vu le nombre total des tags que nous avons utilisé, mais réellement nous sommes affrontés à un grand nombre de tags par ressource, plus les tags sont nombreux plus le nombre composant les descripteurs des ressources augmente.

6. Conclusion

Afin d'améliorer l'approche de filtrage à base du profil utilisateur, nous l'avons dans un premier lieu analysée et extrait ses limites dans le chapitre précédent.

Dans ce chapitre nous avons défini nos motivations, par la suite nous avons présenté les solutions choisies afin de pallier aux limites de l'approche de filtrage basée sur le profil utilisateur, avec la précision des étapes à suivre ainsi que la conception des solutions en appuyant sur un exemple permettant d'illustrer la version améliorée.

Après la conception nous passons à l'étape de la réalisation du système qui est détaillée dans le chapitre suivant.

CHAPITRE IV : Réalisation du système

1. Introduction

Après avoir abordé les solutions à réaliser ainsi que leurs conceptions dans le chapitre précédent nous passons à l'implémentation de l'application qui est la dernière phase dans le processus du développement d'un projet.

Dans ce chapitre nous allons tout d'abord présenter les outils et environnement de développement, nous décrivons par la suite l'application et ses différentes fonctionnalités en l'accompagnant par quelques captures d'écrans.

2. Outils et environnement de développement

Avant de présenter l'application nous tenons à présenter les outils et l'environnement de développement de notre application qui se résume dans ce qui suit :

2.1. Xampp 1.8.2 (X Apache MySQL Perl Php)

XAMPP est un kit d'installation d'Apache qui contient un ensemble de logiciels permettant de mettre en place facilement un serveur, c'est l'outil idéal pour développer et tester des applications PHP, comme il peut être utiliser pour créer et configurer les bases de données écrites dans MySQL **[Dvorski, 07]**.

Il s'agit donc d'une distribution de logiciels libres, Cette distribution se chargera d'installer l'ensemble des outils dont nous pourrions avoir besoin lors de la création d'un site Web.

La version 1.8.2 se compose principalement des outils suivant :

- **Apache 2.4.4**

Apache est une plate-forme serveur web open-source créé et maintenu par un groupe d'étudiants *a he g* [1], permettant à des clients d'accéder à des pages web, c'est-à-dire en réalité des fichiers au format HTML à partir d'un navigateur installé sur leur ordinateur distant [**Fielding et Kaiser, 97**]

Il s'agit d'une application fonctionnant à la base sur les systèmes d'exploitation de type Unix, mais il a désormais été porté sur de nombreux systèmes, dont Microsoft Windows., il existe plusieurs version de apache : 1.3, 2.0…Celle incluse dans xampp est 2.4.4

- **MySql 5.5.32 (My Structured Query Language)**

Un système de gestion de base de données relationnelle libre, qui est très employé sur le Web, souvent en association avec PHP (langage) et Apache (serveur web).

C'est un serveur de base de données qui sauvegarde les données dans des tables séparées plutôt de tout rassemblé dans une seule table, cela améliore la rapidité et la souplesse de l'ensemble. Les tables sont reliées par des relations définie qui rendent possible la combinaison de données entre plusieurs tables en utilisant les requêtes **[MySQL]**. La version de mysql utilisée dans xampp est 5.5.32 sachant qu'il y a plusieurs d'autres versions : 4.0, 4.1, 5.0…

- **PHP 5.4.16 (hypertext preprocessor)**

Php est un langage de scripts, utilisé pour la création des pages web dynamiques, il s'exécute sur un serveur, cela signifie qu'il est exécuté par un serveur avant d'apparaitre en tant que document HTML.

C'est un langage open-source et multiplateforme il fonctionne plus efficacement sur un serveur Apache et sa force réside dans sa compatibilité avec de nombreux types de SGBD (système de gestion de base de données) notamment Mysql **[PHP]**.

- **PhpMyAdmin 4.0.4**

C'est un outil qui facilite l'administration de MySQL sur le web. Il est écrit en PHP et s'appuie sur le serveur http Apache **[PhpMyAdmin]**, il permet notamment de :

- créer / supprimer des bases de données.
- créer / modifier / supprimer des tables ou enregistrements.
- exécuter / modifier / ajouter des requêtes SQL.
- importer et d'exporter des structures ou données d'une base de données Mysql

3. Présentation de l'application

Dans cette partie du chapitre nous allons décrire l'application en illustrant chaque étape avec des captures d'écran permettant d'exploiter au mieux notre système.

3.1. Les interfaces de l'application

Notre application est composée de différentes interfaces que nous allons les détaillées ci-dessous.

3.1.1. Interface principale

Au lancement de notre application un champ d'authentification s'affiche permettant aux utilisateurs d'accéder à leurs comptes, s'ils en possèdent, en introduisant leurs e-mail et mot de passe, sinon ils peuvent s'inscrire en cliquant sur « inscrivez-vous ! »

La figure suivante représente l'interface principale :

Fig 22 : L'interface principale de l'application

3.1.2. Espace utilisateur

Voici les principales fenêtres à la quelles un utilisateur pourra y accéder, mis à part la fenêtre d'authentification qui est affichée au lancement de l'application.

- **Fenêtre d'inscription**

Un utilisateur doit s'inscrire pour pouvoir accéder à notre application, la figure suivant représente les champs à remplir afin de créer un compte personnel.

Fig 23: La fenêtre d'inscription

Donc un utilisateur doit introduire son nom, prénom, sexe (homme/femme), date de naissance, e-mail, ainsi que le mot de passe, pour concevoir son profil.

- **Fenêtre d'accueil**

Une fois que l'utilisateur est authentifié il aura accès à la page d'accueil représentée dans la figure suivante :

Fig 24 : La page d'accueil

L'utilisateur pourra ainsi afficher les ressources en cliquant sur « toutes les ressources », ajouter de nouvelles ressources sur « nouvelle ressource » ou bien rechercher une ressource en cliquant sur « recherche ».

3.1.3. Espace administrateur

L'administrateur comme tout autre utilisateur, doit s'authentifier, comme la montre la figure ci-dessous :

Fig 25 : Fenêtre d'authentification de l'administrateur

- **Fenêtre d'accueil de l'administrateur**

La page d'accueil de l'administrateur est différente de celle de l'utilisateur, comme l'indique la figure suivante :

Fig 26 : Page d'accueil de l'administrateur

L'administrateur peut accéder à la liste des utilisateurs en cliquant sur « Utilisateurs », ainsi que la liste des ressources « Toutes les ressources », il peut également ajouter de nouvelles ressources en appuyant sur « Nouvelle ressource », vérifier les tags suggérés sur

« modération des suggestions » et le bouton « Recherche » permet l'exploration des ressources.

- Visualiser la liste des utilisateurs

Cette liste contient les informations concernant tout les utilisateurs inscrits ainsi que leurs vecteurs d'intérêts et leurs expertises, sachant que l'administrateur a la possibilité de supprimer tout utilisateur qu'il trouve malveillant (fig 27).

Utilisateurs :

informations		vecteurs d'interets	expertise	actions
u2@email.fr Foura Abdelatif	1990-12-24 Genre: m	HIV (15) obesity (13) death (5) symptom (10) Analysis (8)	1.88	[Supprimer]
u3@email.fr foura imad	1995-04-12 Genre: m	artery (7) sick (12) symptom (22) conical (5) cardiomyopathy (2)	3	[Supprimer]
u4@email.fr ihadaden narimen	1990-07-09 Genre: f	Analysis (11) Hypertension (6) artery (4) cardiomyopathy (10) obesity (2)	2	[Supprimer]
u5@email.fr ksentini dalila	1990-10-06 Genre: f	HIV (14) heart failure (13) Pathogens (10) cardiomyopathy (16) Diabetes (8)	2	[Supprimer]

Fig 27 : La liste des utilisateurs inscrits, leurs centres d'intérêts et leurs expertises.

- Ajouter une ressource

L'ajout d'une ressource se fait en introduisant le titre, l'url et le résumé de la ressource ensuite en cliquant sur le bouton « + Ressource » pour valider, comme l'indique la figure suivante :

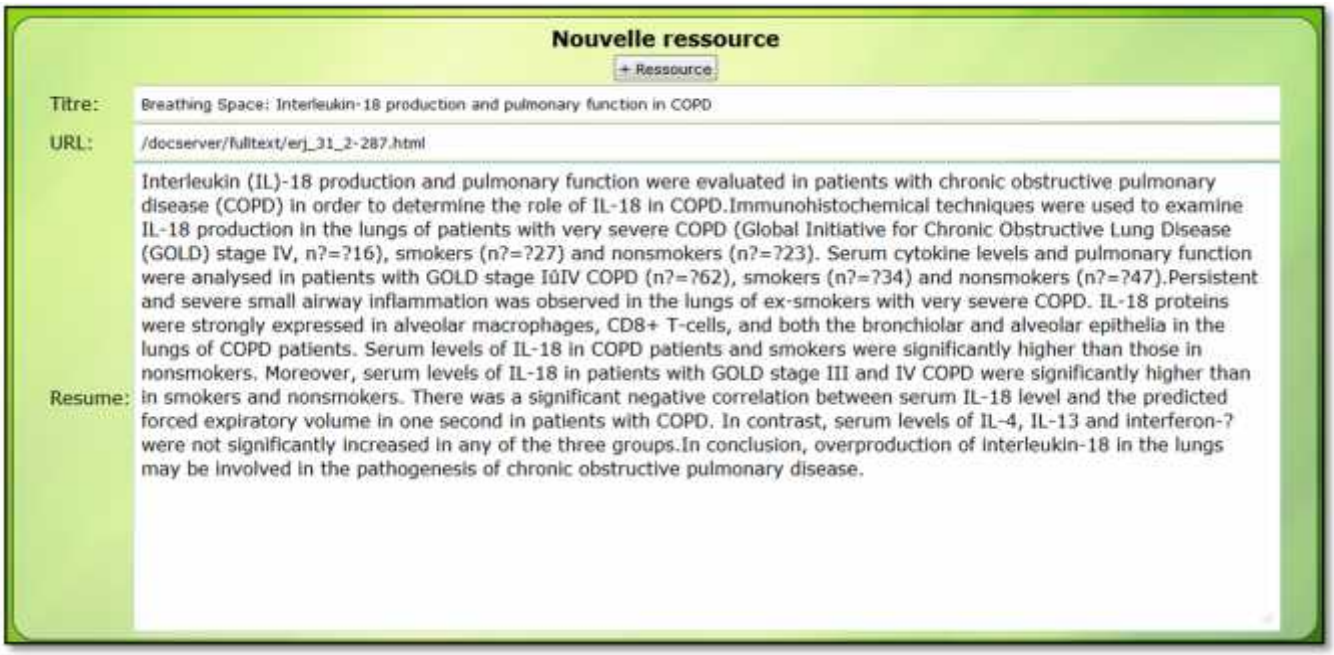


Fig 28 : Ajouter une ressource

- Modération des suggestions

Cette fenêtre permet à l'administrateur de visionner les tags suggérés par les différents utilisateurs.

L'administrateur doit avant tout vérifier le tag, s'il est le descendant ou le synonyme d'un terme de l'ontologie, si c'est le cas il précisera le genre de relation et le terme en considération ensuite il validera la suggestion, sinon il la rejette (fig 29).

Fig 29 : Modération des tags suggérés.

4. Les fonctionnalités offertes par l'application

Notre application met à la disposition de l'utilisateur de diverses fonctionnalités parmi celles-ci nous citons :

4.1. Visualisation des ressources

Les ressources sont réparties sur plusieurs pages, tel que chaque page contient 10 ressources, au total il y a 2609 ressources que l'utilisateur a la possibilité de visionné, voici la figure illustrant une liste de ressource :

Fig 30 : Afficher la liste des ressources

En cliquant sur le titre d'une ressource, cette dernière sera affichée afin d'être à la disposition de l'utilisateur, comme le montre la figure suivante :

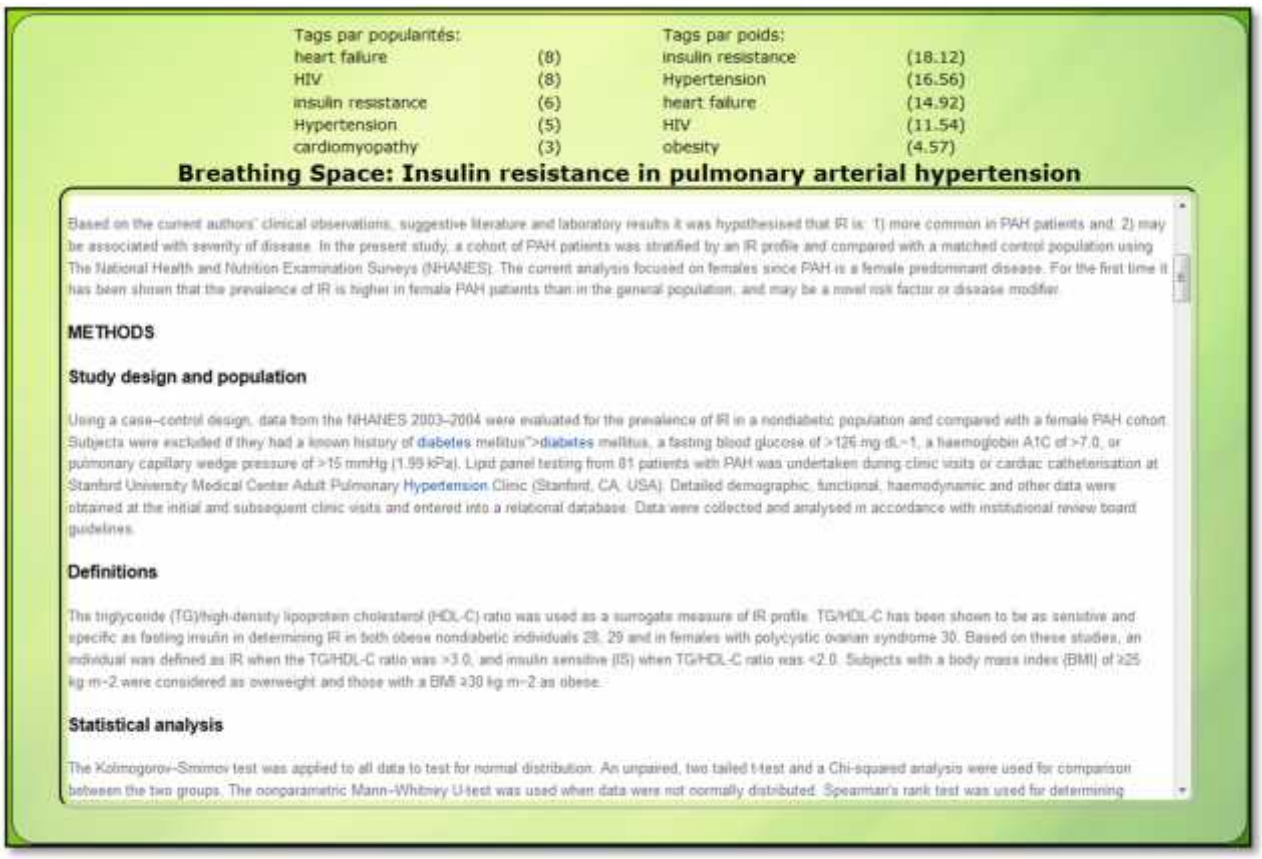


Fig 31 : Visualisation d'une ressource

4.2. Tagguer une ressource

Une fois que l'utilisateur accédera avec son compte, il peut tagguer de multiples ressources, en attribuant au maximum cinq tags pour chacune (fig 32), ainsi l'utilisateur il a la possibilité de modifier l'un de ses tags.

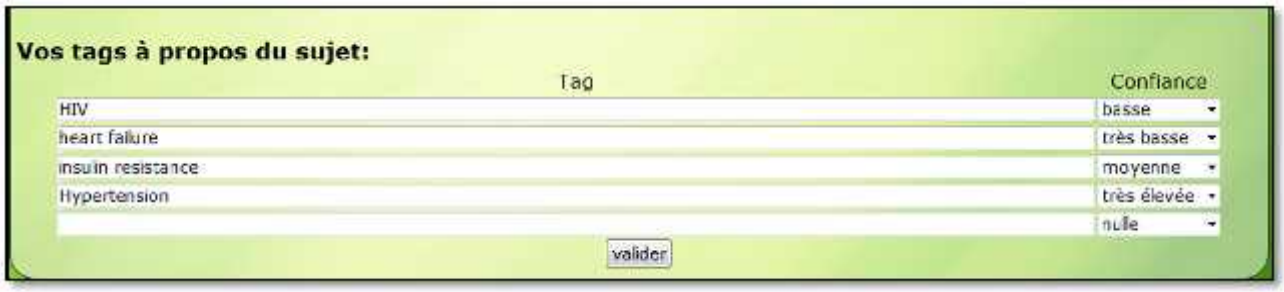


Fig 32 : Tagguer une ressource

Une fois le tag attribué l'utilisateur accordera un degré de confiance par rapport à son tag (fig 33), sachant que nous partons de l'hypothèse que l'utilisateur sera sincère en précisant la confiance.

Fig 33 : Attribution de degré de confiance

La confiance varie comme suit :

- **Nulle** : dans le cas d'un utilisateur qui n'a aucune idée si son tag à un rapport avec la ressource ou non.
- **Très basse** : si l'utilisateur n'est pas du tout sur de son tag.
- **Basse** : en cas ou l'utilisateur pensera qu'il y a peut être une relation entre la ressource et le tag qu'il vient attribuer.
- **Moyenne** : si l'utilisateur connait le domaine de la ressource mais il a quelques doutes par rapport à son tag.
- **Elevée** : c'est-à-dire que l'utilisateur est confiant de son tag mais pas à 100%.
- **Très élevée** : dans le cas ou l'utilisateur est très confiant et il est très sur que le tag qu'il vient d'attribuer concerne exactement la ressource.

4.3. Suggestion des tags

Vu que la propriété principale des systèmes du tagging est la liberté dans le choix des tags, nous avons proposé une fonctionnalité permettant aux utilisateurs de suggérer des tags qui n'appartiennent pas à l'ontologie et qui concerne le domaine d'une quelconque ressource, pour le faire l'utilisateur annotera une ressource avec son nouveau tag au lieu de choisir parmi ceux qui existent dans l'ontologie et clique sur « valider », la fenêtre suivante s'affichera :

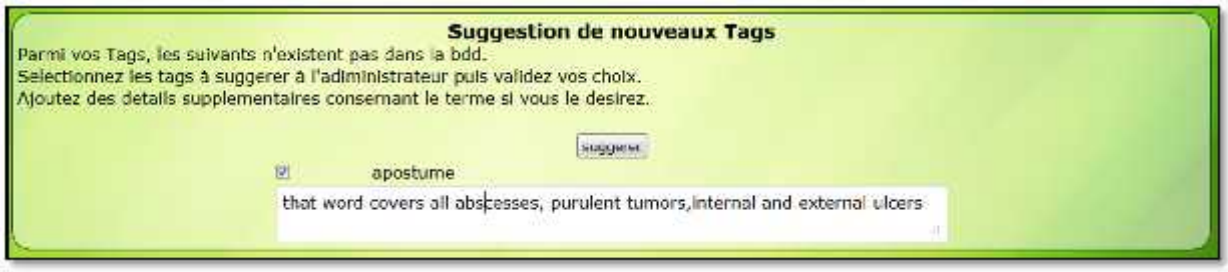


Fig 34 : suggestion de nouveaux tags

L'utilisateur est capable d'introduire des détails supplémentaires concernant son nouveau tag, puis il clique sur « suggérer », en attendant la validation par l'administrateur et ceci pour ne pas toucher le caractère de liberté des choix dont sont connus les systèmes du tagging collaboratif.

4.4. Rechercher une ressource

L'utilisateur a la possibilité de rechercher une ressource en introduisant un tag dans la barre de recherche, les résultats des ressources concernant ce tag sont triés par poids ou popularité, comme le montre la figure suivante :

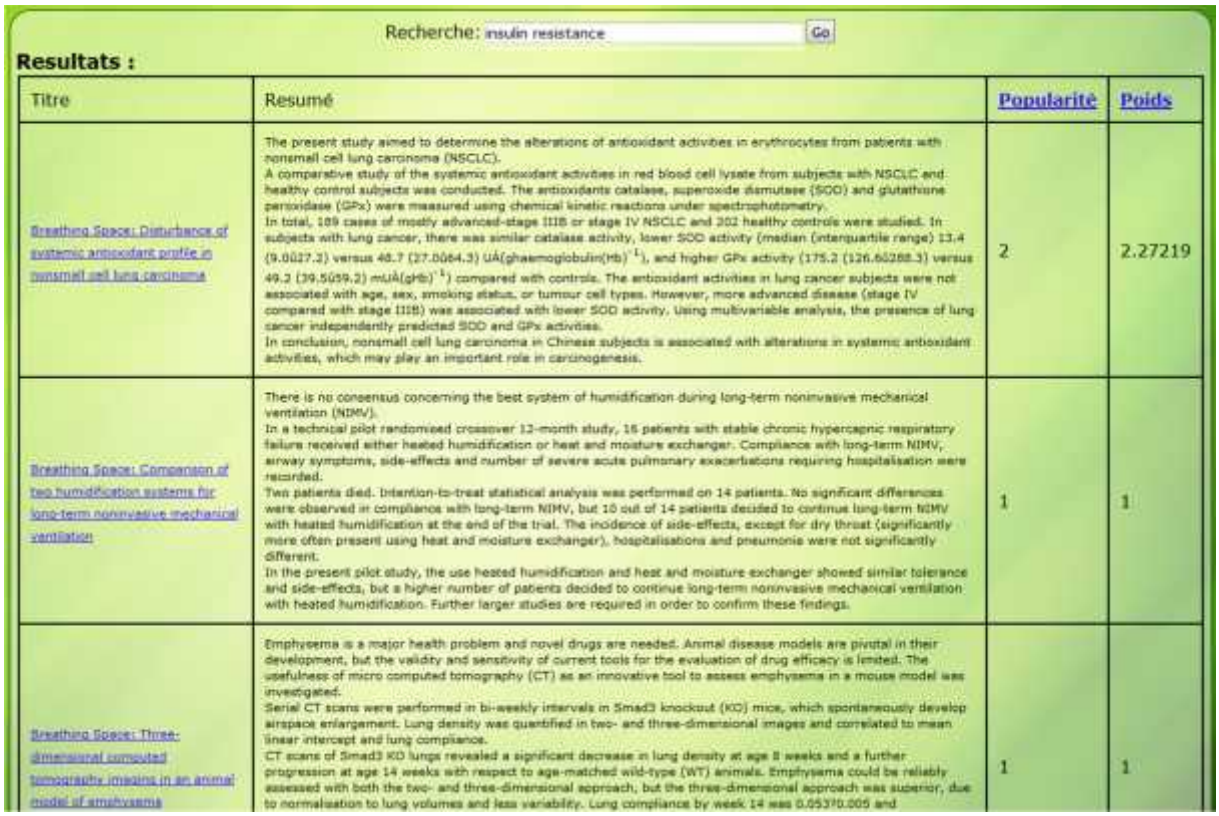


Fig 35 : Recherche des ressources

5. Conclusion

Dans le but d'évaluer notre travail qui consiste à apporter des améliorations à l'approche de filtrage des tags à base du profil utilisateur, nous avons développé un système permettant d'implémenter l'approche ainsi que les extensions que nous avons proposé dans l'intention d'améliorer l'approche de filtrage tout en palliant à ces limites.

Pour cela nous avons décrit en détail notre système tout en présentant les outils et l'environnement de développement, ainsi que les différentes fonctionnalités de l'administrateur et de l'utilisateur en l'illustrant avec des captures d'écran de notre système.

Conclusion Générale

Synthèse

Le terme Web 2.0 est utilisé pour désigner une nouvelle génération des sites internet qui favorisent la collaboration et le partage, ce qui a conduit à l'apparition des systèmes du tagging collaboratif, ces systèmes qui accordent aux utilisateurs la possibilité de tagguer, partager ou même organiser des ressources en ligne.

Nous avons donc défini le tagging en spécifiant ces structures, modèles, propriétés et ces limites, ensuite nous avons présenté l'approche du filtrage des tags à base du profil utilisateur permettant d'améliorer l'accès aux ressources partagées en calculant les poids des tags en incluant le profil utilisateur au lieu de considérer les tags les plus populaires comme étant des descripteurs de ressources.

Notre objectif est donc d'étudier cette approche de filtrage et de l'améliorer, pour cela nous avons tout d'abord présenté l'approche ainsi que ces points forts et ces points faibles, par la suite nous avons proposé des solutions pour pallier à ses limites et pourvoir attribuer les tags les plus adaptés à une ressource.

Résumé de la contribution

Dans le cadre de ce travail nous avons proposé une amélioration pour l'approche de filtrage de tags à base du profil utilisateur **[Kichou, 10]** qui comporte certaines limites.

Tout d'abord nous avons utilisé une ontologie du domaine biomédical, nous avons stemmatisé tous ses termes ainsi que leurs synonymes en utilisant l'algorithme de **[Porter, 80]**, nous avons aussi appliqué la formule de *damerau-levenshtein* **[Bill et Moore, 00]** qui consiste à calculer la distance entre le stemme d'un tag et les stemmes des termes de l'ontologie afin d'identifier le degré de rapprochement entre eux.

Une fois la distance calculée nous suggérons à l'utilisateur tous les termes ainsi que leurs synonymes identiques à son tag, de cette façon nous éliminerons le problème de la variation d'écriture des tags ainsi que l'ambigüité.

Notre travail a été évalué en utilisant une collection de ressources extraites de *http://respiratory.publishingtechnology.com/*, il n'est qu'une simple extension dans le domaine du tagging collaboratif et son exploitation sera importante particulièrement dans la recherche d'informations.

Perspectives

Les perspectives envisageables pour notre travail sont :

- Evaluer l’approche sur une collection plus importante, que sa soit en nombre de tags, de ressources ou d’utilisateurs.
- Etendre l’application pour d’autres domaines.
- Permettre l’utilisation du multilinguisme (Anglais et Français).

Bibliographie

[Attwood et al, 11]

TK. Attwood, A. Gisel, NE. Eriksson and E. Bongcam Rudloff «*Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective* ». Bioinformatics-Trends and Methodologies, 2011.

[Brill et Moore, 00]

E. Brill, R.C. Moore **«***An improved error model for noisy channel spelling correction*». In Proceeding of the 38th annual Meeting on Association for computational linguistics, p 286-293, Stroudsburg, USA, 2000.

[Cattuto et al 07]

C. Cattuto, V. Loretto and L. Pietronero «*Semiotic Dynamics and Collaborative Tagging*». In Proceedings of National Academy of Sciences of the USA, 104(5): 1461-1464, originally published online Jan 23, 2007.

[Cayzer et Michlmayr, 09]

S. Cayzer, E. Michlmayr «*Adaptive User Profiles*». In *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*, ISBN-13 :9781605663067, p 65-87, 2009.

[Chelliah et al, 13]

V. Chelliah, C. Laibe and N. Le Novère «*BioModels Database: A Repository of Mathematical Models of Biological Processes*». In silico systems biology, 2013.

[Commontag]

Commontag – QuickStartGuide «En ligne». [http://commontag.org/Home] (Consulté le 29/03/2013).

[Courtot et al, 07]

M. Courtot, N. Le Norvége and C. Laibe «*Adding Semantics in kinetics models of biochemical pathways*». In proceedings of the 2nd international symposium on experimental standard conditions of enzyme characterizations, 2007.

[Deuff, 06]

O.Le Deuff «*Folksonomies : Les usagers indexent le web*». Bulletin-Bibliothèques de France ISSN : 1292-8399, 2006.

[Dvorski, 07]

D.D. Dvorski «*Installing, Configuring and Developing with XAMPP*». Skills Canada, 2007.

[**Fielding et Kaiser, 97**]

R.T. Fielding, G. Kaiser *«The Apache Http Server Porject»*. IEEE Internet Computing 1(4): 88-90, USA, 1997.

[Furnas et al, 87]

G.W. Furnas., T.K. Landauer, L.M. Gomez and S.T. Dumais, «*The vocabulary problem in human-system communication*». Communications of the ACM, 30(11): 964-971, New York, USA, 1987.

[Fu, 08]

W.T. Fu «*The microstructures of social tagging: a rational model*». In Proceedings of the ACM conference on Computer supported cooperative work, p 229-238, USA, 2008.

[Garcia-castro et al, 09]

L.J. Garcia-castro, M. Hepp, A. Garcia «*Tags4Tags: Using Tagging to Consolidate Tags*». In Proceedings of the 20th international conference on Database and expert systems applications, p 619-628, Berlin, 2009.

[Gerald, 02]

J.K. Gerald, T.M. Maybury «*Information storage and retrieval systems Theory and Implementation*», Second Edition, KLUWER ACADEMIC PUBLISHERS, 2002.

[Golder et Huberman, 05]

S.A. Golder, B.A. Huberman «*The Structure of Collaborative Tagging Systems* », CoRR abs/cs/0508082, 2005.

[Golder et Huberman, 06]

S.A. Golder, B.A. Huberman «*Usage patterns of collaborative tagging systems* ». Journal of Information Science, 32(2): 198-208, USA, 2006.

[Gruber, 07]

T.R. Gruber *«Ontology of folksonomy: A mash-up of apples and oranges*». Int. Journal on Semantic Web and Information Systems, 2007.

[Guy et Tonkin, 06]

M. Guy, E. Tonkin *«Folksonomies: Tidying up tags?* ». D-Lib Magazine, 12(1), http://www.dlib.org/dlib/january06/guy/01guy.html, ISSN : 1082-9873, 2006.

[Halpin et al, 07]

H. Halpin, V. Robu and H. Shepherd *«The complex dynamics of collaborative tagging*». In Proceedings of the 16th international conference on World Wide Web, p 211-220, USA, 2007.

[Hastings et al, 08]

J. Hastings, K. Degtyarenko, P. De Matos, M. Ennis, , M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj and M. Ashburner *«ChEBI: a database and ontology for chemical entities of biological interest»*. Nucleic acid research, 2008.

[Hedstrom, 05]

P Hedstrom *«Dissecting the social: On the principales of analytical sociology*». Cambridge University Press: Cambridge, 2005.

[Kichou, 10]

S.Kichou *«Tagging collaboratif et filtrage de tags à base du profil utilisateur*». Thèse de Magistère en Informatique, Université Abderrahmane Mira De Bejaia, 2010.

[Kichou, 11]

S. Kichou, Y. Amghar, H. Mellah «*Approche de filtrage de tags à base du profil utilisateur*». First International Workshop on Semantic and Collaborative Technologies for the Web, Bucharest, Romania, juin 2011.

[Levenshtein, 96]

V.I Levenshtein. **«** *Binary codes capable of correcting deletions, insertions and reversals* ». Soviet Physics Doklady, 10(8):707–710, 1996.

[Lovins, 68]

J.B. Lovins «*Development of a stemming algorithm* ». Mechanical translation and computational linguistics, 11: 22-31, 1968.

[Marlow et al, 06]

C. Marlow, M. Naaman, D. Boyd and M. Davis « *tagging paper, taxonomy, flickr, academic article, to read* ». In Collaborative Web tagging workshop at WWW'06, Edinburgh, UK, 2006.

[MySQL]

MySQL – Documentation MySQL server, *What is MySQL?* « En ligne », [http://dev.mysql.com/doc/refman/4.1/en/what-is-mysql.html] (Consulté le 21/07/2013).

[Namer et Baud, 07]

F. NAMER, R. BAUD «*Defining and relating biomedical terms: towards a cross-language morphosemantics-based system*». Int. Journal of Medical Informatics, 76(2-3):226–233, 2007.

[Paice, 90]

C.D. Paice «*Another stemmer*». ACM SIGIR forum, 24(3): 56-61, 1990.

[Passant, 07]

A. Passant «*Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs*». In Proceedings of international Conference on Weblogs and Social Media, 2007.

[Passant et Laublet, 08]

A. Passant, P. Laublet «*Meaning Of A Tag : A collaborative approach to bridge the gap between tagging and linked data* ». In proceedings of the linked data on the Web, 2008.

[Paternostre et al, 02]

M. Paternostre, P. Francq, J. Lamoral, D. Wartel and M. Searens *«Carry, un algorithme de désuffixation pour le français* ». Generic Analyser and Listener for Indexed and Linguistics Entities of Information, 2002.

[PHP]

PHP – *«What is PHP?* » [En ligne]. http://www.php.net/manual/en/intro-whatis.php, (Consulté le 02/08/2013).

[PhpMyAdmin]

PhpMyAdmin *«About PhpMyAdmin»*, [En ligne]. http://www.phpmyadmin.net/home_page/index.php/, (Consulté le 20/08/2013).

[Porter, 80]

M.E. Porter *«An algorithm for suffix stripping»*. Program: Electronic library and information systems, 14(3): 130-137, 1980.

[Roxin et Bernard, 07]

V. Roxin, Y. Bernard *« Etiquetage collaboratif et nuages de mots : quels apports pour les sites marchands* ? ». Actes de la 6éme journée Nantaise de recherche sur le e-marketing, Nantes, 2007.

[Sen et al, 06]

S. Sen, S.K. Lam, A.M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. Maxwell Harper and J. Riedl *«tagging, communities, vocabulary, evolution»*. In Proceedings of the 2006 20th anniversary conference on Computer supported, CSCW : p 181-190, New York, USA, 2006.

[Sinclair et Cardew-Hall, 08]

J. Sinclair, M. Cardew-Hall ***«****The folksonomy tag cloud: when is it useful?* ». Journal of information science, 34(1): 15-29, USA, 2008.

[Specia et Motta, 07]

L. Specia, E. Motta *«Integrating folksonomies with the semantic web* ». In 4th European Semantic Web Conference, 2007.

[**Vander Wal ,07]**

T. Vander Wal *« Folksonomy Coinage and Definition »*, 2007. [En ligne] http://vanderwal.net/folksonomy.html. (Consulté le 16/03/2013).

[**Vander Wal ,05]**

T. Vander Wal *« Explaining and Showing Broad and Narrow Folksonomies »*, 2005. [En ligne]. http://www.vanderwal.net/random/entrysel.php?blog=1635. (Consulté le 16/03/2013).

[Wang, 10]

J. Wang, M. Clements, J. Yang, A.P. De Vries, J.T. Reinders *«Personnalization of tagging systems »*. Information Processing and Management, 10

[Wasserman, 94]

S. Wasserman *« Social Network Analysis: Methods and Applications»*. Cambridge, UK: Cambridge University Press, 1994.

[Wetzker et al, 10]

R. Wetzker, C. Zimmermann, C. Bauckhage and S. Albayrak *« I tag, you tag: translating tags for advanced user models»*. In Proceedings of the third ACM international conference on Web search and data mining, WSDM: 71-80, New York, USA, 2010.

Annexe A : Exemple des systèmes de tagging

Durant ces dernières années les sites du tagging sont de plus en plus fréquents, **[Marlow, 06]** dans ses travaux a choisi douze exemples qu'elle estime divers dans leurs architecture et leurs types de ressources, parmi ces sites :

- **Flickr (http://www.flickr.com/)**

Flickr est l'un des meilleurs sites de gestion et de partage des photos en ligne, il héberge plus de 5 milliards de photos du monde entier (fig)

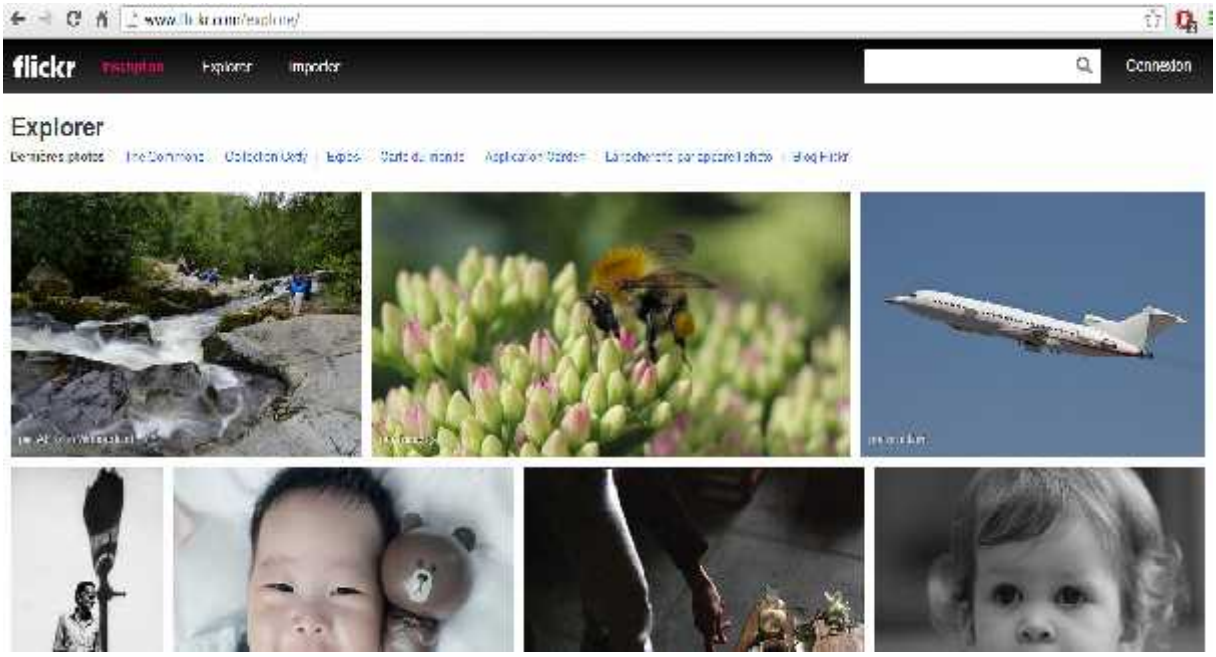


Fig 1 : page d'accueil de flickr

- **Del.icio.us (http://delicious.com/)**

Le plus connu des sites de bookmarking (marque-page) permettant aux utilisateurs de partager, sauvegarder et de tagguer des pages web, la figure suivante montre le résultat de recherche du tag java sur delicious :

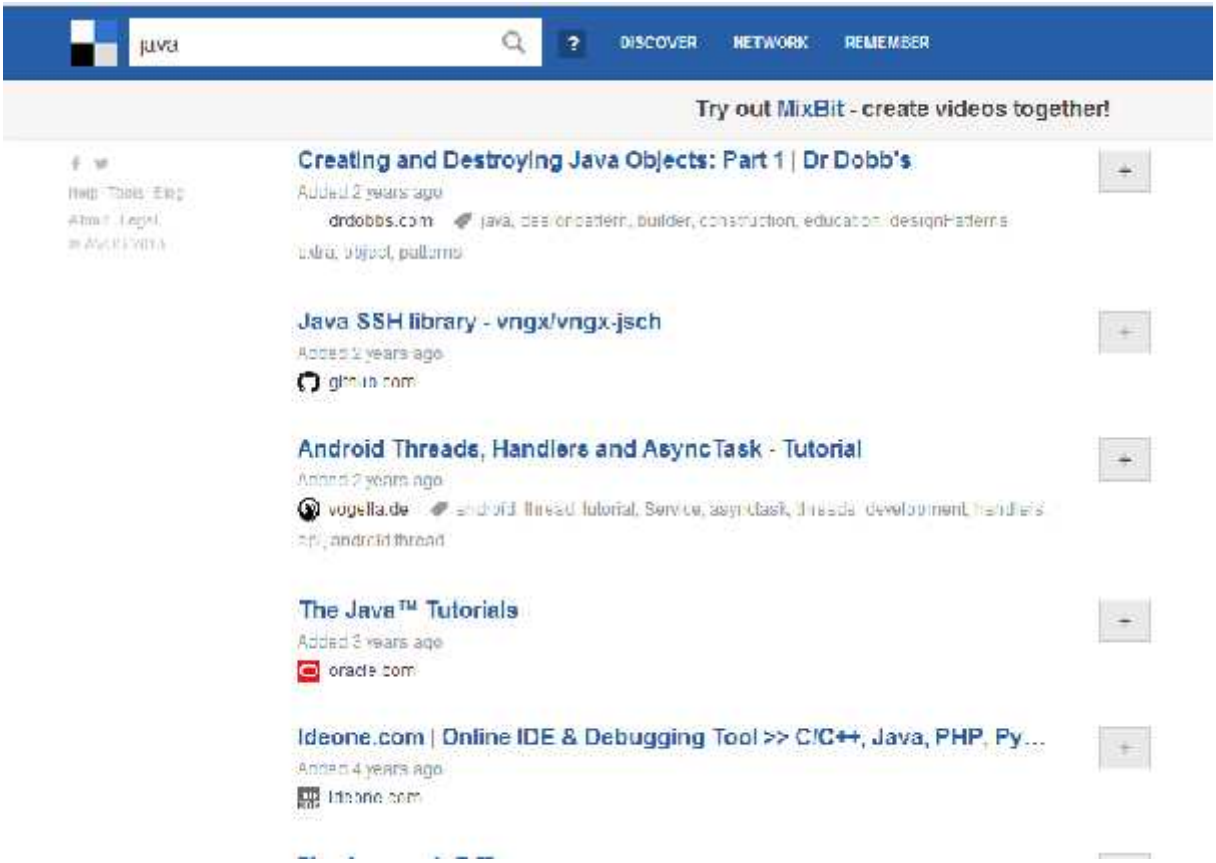


Fig 2 : résultat de recherche du tag java sur delicious

- **citeUlike (http://www.citeulike.org/)**

C'est l'un des plus populaires sites permettant la gestion des références scolaires tout en aidant les utilisateurs à stocker et découvrir des références académiques :

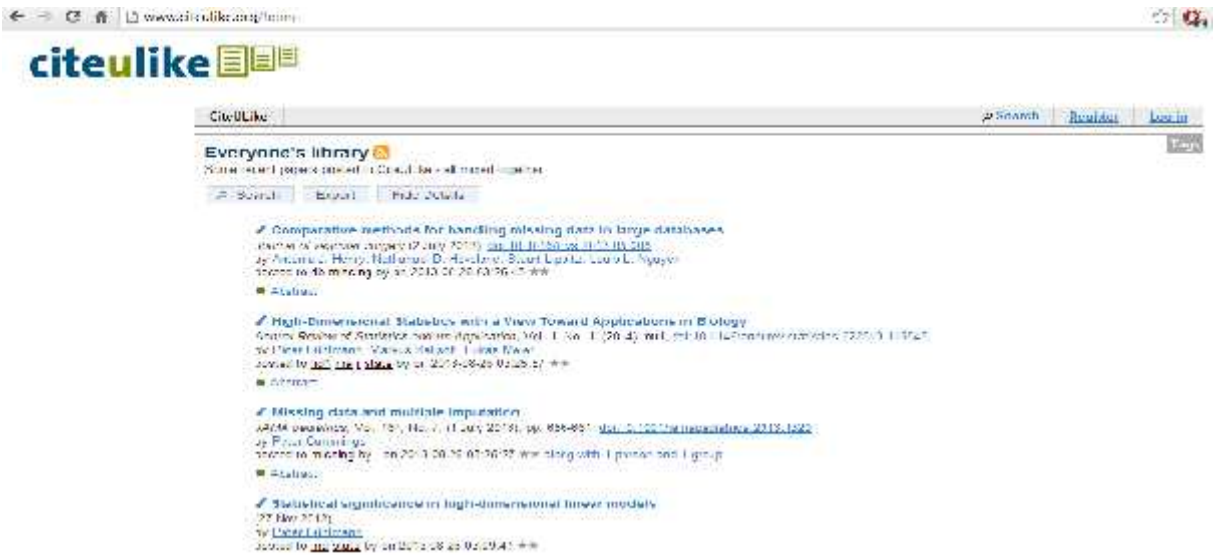


Fig 3 : page d'acceuil de citeUlike

- **Technorati (http://technorati.com/)**

Est un moteur de recherche en temps réel, il cherche seulement à travers les blogs, permettant aussi aux utilisateurs d'attribuer des tags :

Fig 4 : liste des tags les plus populaires sur technorati

- **Youtube (http://youtube.com/)**

C'est un site de partage de vidéo, qui permet aux utilisateurs de partager, tagguer ou visualiser des vidéos en ligne

- **Yahoo MyWeb 2.0 (http://myweb.yahoo.com/)**

Permet aux utilisateurs d'importer des ressources et de les organiser dans des dossiers toute en incluant un réseau social.

- **ESP Game (http://espgame.com/)**

Un jeu de tagging sur internet, permettant à deux utilisateurs choisis aléatoirement de tagguer une ressource en générale une image, à condition de choisir le même tag afin de passer à l'image suivante.

- **Last fm (http://last.fm/)**

Le plus grand catalogue de musique en ligne permettant aux utilisateurs de tagguer les artistes, les albums et les chansons.

- **Yahoo ! Podcasts (http://podcasts.yahoo.com/)**

Permettant de chercher des podcasts (diffusion des contenus audio) et les tagguer.

- **Odeo (http://odeo.com/)**

Un service de podcasting permettant aux utilisateurs de cherche, écouter et tagguer des podcasts.

- **LiveJournal (http://livejournal.com/)**

Un service pour les journaux et les blogs qui favorise les échanges communautaires.

- **Upcoming(http://upcoming.org)**

Un site qui autorise les utilisateurs d'introduire des événements (pièce de théâtre, concert..) et de les tagguer.

Annexe B : L’ontologie biomédicale

- ## Introduction

Nous avons utilisé dans notre travail une ontologie de domaine biomédicale, afin de pallier aux limites d’ambigüité et de variations d’écritures.

Dans cette annexe nous allons décrire l’ontologie et nous allons citer quelques ontologies biomédicales et enfin nous parlerons des relations sémantiques entre les termes biomédicaux.

- ## Description

L’ontologie biomédicale est une base de données contenant plusieurs sous ontologies des différentes spécialités médicale.

Cette ontologie a été développée par l’institut européen bioinformatique (EBI) qui fait partie de laboratoire européen de biologie moléculaire (EMBL) qui est une organisation intergouvernementale constituée de 20 états et dirigée par le directeur général, le professeur *Lain Mattaj* nommé par le conseil.

En 1992 le conseil EMBL à voté pour la création de l’institut européen bioinformatique (EBI), la vision pour la création de cet institut est rapidement survenu et il est devenu pleinement opérationnel en 1995, il est composé de trois grandes divisions : la recherche, l’industrie et les services.

La mission d’EBI était de s'assurer que les données issues de la biologie moléculaire et la recherche sur le génome a été placé dans le domaine public et est accessible librement pour toute la communauté scientifique dans l’intention de favorisé les progrès scientifique.

Aujourd’hui l’institut construit, entretien et diffuse des bases de données et des services d’informations pertinentes à la biologie moléculaire, la génétique, la médecine et l’agriculture **[Attwood et al, 11].**

L’ontologie biomédicale est constituée exactement de 86 ontologies et 2034847 termes.

- **Quelques ontologies biomédicales**

- SBO (System Biology Ontology) : est un ensemble de vocabulaires contrôlés, comportant les termes couramment utilisés en biologie des systèmes, les termes de cette ontologie sont liés par la relation d’héritage « is a » **[Courtot et al, 07]**.

- BioModels : la base de données BioModels sert comme de référentiel fiable de modèles computationnels des processus biologiques **[Chelliah et al, 13]**.

- ChEBI (Chemical entities of biological interest): une base de données des entités moléculaires, plus précisément les composants chimiques de petite taille **[Hastings et al, 08]**.

- **Les relations sémantiques entre les termes biomédicaux**

Selon **[Namer et Baud, 07]** il existe trois types de relations sémantiques entre les termes biomédicaux :

- L’hypéronymie : terme général dont le sens inclut celui d’un ou plusieurs autres termes, si le terme B et hyperonyme de A c'est-à-dire que A fait partie de B.

- La synonymie : deux termes ayant le même sens.

- La Méronymie : un terme A est méronyme de B c'est-à-dire que A fait partie de B.