République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderrahmane Mira de Bejaia
Faculté des Sciences Exactes
Département d'Informatique

ECOLE DOCTORALE RESEAUX ET SYSTEMES DISTRIBUES

Mémoire de Magistère

En Informatique

Option : Réseaux et Systèmes Distribués

Thème

Tagging Collaboratif et Filtrage de Tags à base du Profil Utilisateur

---

Présenté par

Mme DAHAK-KICHOU Saida

Devant le jury composé de :

Président	TARI Abdelkamel	M.C.A	Université de Bejaia
Rapporteur	AMGHAR Youssef	Professeur	INSA Lyon France
Examinateur	DRIAS Habiba	Professeur	USTHB Alger
Examinateur	MAHDAOUI Latifa	M.C.A	USTHB Alger
Invitée	MELLAH Hakima	Chargée de recherche	CERIST Alger

Promotion : 2008/2009

Remerciements

Merci au bon Dieu, le tout Puissant.

Je tiens à remercier vivement mon directeur de thèse Mr Amghar Youssef, professeur à l'INSA, pour ses corrections, conseils et orientations. Merci d'avoir accepté de m'encadrer et me guider dans le monde si vaste de la recherche. Qu'il trouve en ce mémoire l'expression de mon profond respect.

Je remercie tout particulièrement Mme Mellah Hakima, chargée de recherche au CERIST, codirecteur de thèse, de m'avoir orienté, corrigé mon travail, soutenu et encouragé. Merci pour sa disponibilité et sa gentillesse sans égale. Je ne la remercierai jamais assez, qu'elle trouve en ce mémoire l'expression de ma profonde gratitude et mon respect infini.

Mes vifs remerciements et respects s'adressent aux membres du jury Mr Tari Abdelkamel, maitre de conférence à l'université de Bejaia, Mme Drias Habiba, professeur à l'USTHB et Mme Mahdaoui Latifa, maitre de conférence à L'USTHB pour m'avoir fait l'honneur d'accepter de juger ce modeste travail.

Je tiens également à remercier Mr Badache Nadjib, professeur à l'USTHB et Directeur du CERIST, de m'avoir donné la chance d'entrer dans le monde de la recherche.

Mes sincères remerciements vont également à Mr Meziane Abdelkrim, chargé de recherche et responsable de la division systèmes d'information et systèmes multimédias pour ses encouragements et sa compréhension.

Je remercie infiniment Mr Dahak Fouad, maitre assistant à l'école supérieure d'informatique pour son aide très précieuse, ainsi que Mr Azouaou Faiçal, maitre de conférence pour ses orientations et conseils.

Mes chaleureux remerciements vont à mes très chers parents qui ne cessent de croire en moi et m'encourager. Merci pour toute ma famille, grands parents, oncles et tantes, cousins et cousines, et toute ma belle famille.

Mes chaleureux remerciements vont également à mes amies du CERIST Nour El Houda, Linda, Lamia, Fouzia, Safia, Amel et Leila pour leurs soutiens et encouragements. Merci à tous mes collègues de la division Imen, Badiaa, F.Zohra, Lamia, Djalila, Lydia, Salem, Kamal et Noureddine pour leurs encouragements.

Je remercie mes collègues de la promotion, Lamia, Sahar, Dalila, Faiza, Nabil, Madjid et Rafik pour la bonne ambiance vécue tout au long d'une année de cette belle expérience.

Merci à Mme Zaidi et tous les collègues du service formation pour leur disponibilité.

Je remercie mes amis Malika.B, Malika.A, Radia, Sekoura, Soraya et Fazia pour le soutien moral tant apprécié.

Fouad, merci encore une fois, je ne te remercierai jamais assez pour tout ce que tu fais pour moi, merci d'être à mes côtés dans les moments difficiles, sans toi je n'aurai jamais pu tenir jusqu'à la fin.

Tout simplement Merci !

A mes très chers parents Rahima et Abdelmadjid

A mon mari Fouad

A mes deux adorables petits anges Anis et Yacine

A ma chère sœur Samia

A mes frères Mohamed et Azedine

A mes beaux parents Said et Wardia

A jida Adidi, jeddi Omar et ami Mustapha

A toute ma famille et belle famille

RESUME

Le ‘Tagging collaboratif’ ne cesse de gagner une popularité sur le web 2.0, cette nouvelle génération du web qui fait de l’utilisateur un lecteur-rédacteur. Le ‘Tagging’ est un des moyens permettant à l’utilisateur de s’exprimer librement via des ajouts d’étiquettes appelées ‘Tags’ à des ressources partagées. L’un des problèmes rencontrés dans les systèmes du Tagging actuels est la définition des tags les plus appropriés pour une ressource. Les tags sont généralement classés par ordre de popularité tel que del-icio-us. Or la popularité du tag ne reflète pas toujours son importance et sa représentativité vis-à-vis de la ressource à laquelle il est associé. Partant des hypothèses qu’un même tag pour une ressource peut prendre des significations différentes selon les utilisateurs, et un tag issu d’un utilisateur connaisseur serait plus important qu’un tag issu d’un utilisateur novice, nous proposons une approche de pondération des tags d’une ressource à base du profil utilisateur. Pour ceci nous définissons un modèle utilisateur permettant son intégration dans le calcul du poids d’un tag ainsi qu’une formule de calcul de ce dernier à base de trois facteurs concernant l’utilisateur à savoir, le degré de rapprochement entre ses centres d’intérêts et le domaine de la ressource, son expertise et son estimation personnelle vis-à-vis des tags qu’il associe à la ressource. Un descripteur de ressource contenant les meilleurs tags est ainsi créé.

Mots-clés

Annotation, Tagging collaboratif, profil utilisateur, recherche d’informations.

ABSTRACT

The 'Collaborative Tagging' is gaining popularity on Web 2.0, this new generation of Web which makes user reader/writer. The 'Tagging' is a means for users to express themselves freely through additions of label called 'Tags' to shared resources. One of the problems encountered in current tagging systems is to define the most appropriate tag for a resource. Tags are typically listed in order of popularity, as del-icio-us. But the popularity of the tag does not always reflect its importance and representativeness for the resource to which it is associated. Starting from the assumptions that the same tag for a resource can take different meanings for different users, and a tag from a knowledgeable user would be more important than a tag from a novice user, we propose an approach for weighting resource tags based on user profile. For this we define a user model for its integration in calculating the weight of a tag and a formula for calculating it based on three factors namely the user, the degree of approximation between its centers interest and the field of resource, expertise and personal assessment for tags associated to the resource. A resource descriptor containing the best tags is created.

Keywords

Annotation, Collaborative Tagging, user profile, information retrieval.

مـلـخـص

اكتسب التوسيم التعاوني شعبية كبيرة في الواب 2.0 ، النسخة الجديدة من الشبكة العنكبوتية التي جعلت المستخدم قارئ و محرر في آن واحد. التوسيم أحد الوسائل التي تسمح للمستخدم أن يعبر عن رأيه بكل حرية و ذلك بإضافة علامات بشكل كلمات تسمى 'Tag' لموارد مشتركة على الشبكة. من بين نقائص أنظمة التوسيم التعاوني الحالية التعريف بالعلامات المناسبة أكثر لمورد ما. هذه العلامات مرتبة حسب شعبيتها (عدد مرات ذكرها) مثل النظام Delicious. لكن شعبية العلامة لا تعكس دائما أهميتها و تمثيلها بالنسبة للمورد المنسوبة إليه. بافتراض أنه يمكن لعلامة ما أن تحمل معاني مختلفة حسب المستخدمين، و علامة مقترحة من مستخدم خبير أكثر أهمية من علامة مقترحة من مستخدم مبتدأ، نقترح نظاما لترجيح العلامات مستندا على مواصفات المستخدم. لذلك نقترح نموذجا لمواصفات المستخدم، يسمح هذا النموذج بإدخال المواصفات في عملية ترجيح العلامات. و صيغة لحساب وزن كل علامة مستندا على ثلاثة عوامل: درجة التقارب بين المورد و مراكز اهتمام المستخدم، خبرة المستخدم و تقييمه الشخصي لعلاماته. بذلك نكوّن مجموعة أحسن العلامات.

الكلمـــات المفتـــاحية: الملاحظات، التوسيم التعاوني، مواصفات المستخدم، استرجاع المعلومات.

Table des Matières

Introduction Générale

Première Partie:Etat de l'Art

Chapitre I : Les Annotations

Chapitre II : Le Tagging Collaboratif

Chapitre III : Le Profil Utilisateur

Liste des Figures

Liste des Tableaux

Liste des Formules

Introduction Générale

Croire ou ne pas croire, cela n'a aucune importance.

Ce qui est intéressant, c'est de se poser de plus en plus de questions.

(Edmond Wells, Encyclopédie du Savoir Relatif et Absolu, Tome IV).

Introduction

De nos jours, nous vivons dans une société moderne où la place est au web et ses technologies. Avec l'avènement de ce qu'appelle les uns '*Nouveau web*', le rôle des utilisateurs s'est sensiblement modifié. Ce *Nouveau web*, appelé Web 2.0 ou Web collaboratif ou encore participatif symbolise le Web interactif où l'utilisateur n'est plus seulement un consommateur mais aussi un producteur d'information. Cette seconde génération d'internet est basée sur des services permettant aux utilisateurs de collaborer et partager des ressources en ligne. A l'origine du terme Web 2.0 Tim O'Reilly, dans son article fondateur « What is Web 2.0 ? » redéfinit le web non comme une collection de sites Web mais comme une plate-forme, un socle d'échanges entre les utilisateurs.

Suite à cette nouvelle arrivée, de nouvelles fonctionnalités ont vu le jour, le 'Tagging collaboratif' (*Collaborative Tagging*), est l'un des moyens permettant aux utilisateurs d'organiser et partager des ressources en ligne (documents textes, images, vidéos…) via des ajouts d'étiquettes appelées tags pour décrire le contenu de ces ressources.

Contexte du travail et problématique

Le travail que nous présentons dans ce mémoire rentre dans le cadre de l'amélioration des fonctionnalités des systèmes du Tagging collaboratif actuels. Entre autres, ces systèmes permettent aux utilisateurs d'associer, de manière collaborative, des tags à des ressources partagées, ces tags sont susceptibles de décrire au mieux les contenus auxquels ils sont associés. Que ce soit pour des objectifs de recherche ou de classification, l'ajout de tags permet une exploitation de ces contenus.

Pour que cette exploitation soit meilleure, les tags associés à une ressource doivent représenter le plus efficacement possible le contenu de celle-ci.

Entre avantages et inconvénients, le Tagging semble susciter beaucoup d'intérêt, une des voies que nous explorons à travers notre travail est celle de la sélection des tags les plus appropriés pour une ressource, en se basant sur de nouveaux critères.

Les systèmes du Tagging actuels, classent les tags par popularité, celle-ci est définie par le nombre de fois que le tag est cité pour une ressource donnée. D'une part et selon **[Cayzer, 09]**, il est très fréquent qu'un utilisateur répète les mêmes tags déjà associés à la ressource ce qui peut rendre ce tag répété populaire sans qu'il soit pour autant pertinent. Les besoins en information de l'utilisateur ont peu de chance d'être satisfaits **[Wang, 09]**. D'une

autre part, ces tags sont associés par des utilisateurs de niveaux de connaissances et d'expertise différents. Nous estimons donc que la popularité ainsi calculée n'est pas suffisante pour dire que ce tag populaire est représentatif pour une ressource donnée et qu'il faut d'autres critères pour décider de l'adéquation du tag.

L'idée donc de notre travail est d'exploiter cette différence entre les utilisateurs pour filtrer les meilleurs tags d'une ressource.

Contribution

Partant des hypothèses qu'un même tag pour une ressource peut prendre des significations différentes selon les utilisateurs, et un tag issu d'un utilisateur connaisseur serait plus important qu'un tag issu d'un utilisateur novice. Nous proposons une approche de filtrage de tags à base du profil utilisateur.

Notre contribution dans le cadre de ce travail, se résume essentiellement dans l'approche que nous proposons. Nous situons cette contribution à quatre niveaux :

1- *Le modèle utilisateur :* afin de sa prise en considération pour la sélection des meilleurs tags, nous avons défini un modèle utilisateur composé de trois (03) dimensions : personnelle, centres d'intérêt et expertise.

2- *La construction de la dimension centres d'intérêts :* pour la construction de cette dimension, nous avons proposé de combiner deux approches existantes, l'approche naïve et l'approche par cooccurrence.

3- *La construction de la dimension expertise :* pour définir l'expertise d'un utilisateur dans un domaine donné, nous avons proposé une formule basée sur les profondeurs des tags, utilisés par cet utilisateur, dans une ontologie.

4- *La formule de pondération des tags :* Enfin pour pouvoir classer les tags avec notre approche, nous proposons une formule de pondération d'un tag basée sur trois facteurs : la distance entre l'intérêt de l'utilisateur ayant associé ce tag, son expertise et sa confiance vis-à-vis du tag, un autre facteur que nous introduisons.

D'un autre côté, étant confronté au problème d'absence de benchmarks et de plateformes de test dans le domaine du Tagging collaboratif, nous proposons une méthode de test et d'évaluation en nous projetant dans un système de recherche d'information. Il s'agit donc d'évaluer l'approche à travers un domaine d'application. Cette évaluation a permis de montrer clairement l'apport de notre approche.

Organisation du mémoire

Ce mémoire est organisé comme suit :

Première partie : Etat de l'art

C'est une synthèse concernant un ensemble de notions liés au contexte de notre travail. Le Tagging est avant tout un type d'annotation, nous abordons donc en premier lieu la notion d'annotation dans le chapitre I, puis le Tagging collaboratif dans le chapitre II et notre critère de filtrage, le profil utilisateur sera l'objet du chapitre III.

- *Chapitre I : Les annotations :* Dans ce chapitre, nous donnons un ensemble de définitions d'annotations, trouvées dans la littérature. Nous citons également la structure d'une annotation, ses modèles de représentation, ses objectifs selon plusieurs auteurs, ses catégories et nous définissons brièvement ce qu'est l'annotation sémantique. Un ensemble d'outils d'annotations sont présentés dans l'annexe A.
- *Chapitre II : Le Tagging collaboratif :* Dans ce chapitre il est question de définir le Tagging, les différentes notions qu'on lui associe en particulier la folksonomie, ses avantages et inconvénients. Nous enchainons par les structures existantes d'une action du Tagging, puis nous citons les principales propriétés d'un système de Tagging collaboratif. Des exemples de ces systèmes sont également cités dans ce chapitre ainsi qu'un ensemble de limites. Nous élaborons une petite comparaison entre la notion d'annotation et celle du Tagging. Nous survolons un ensemble de travaux de recherche réalisés concernant le Tagging et les folksonomies.
- *Chapitre III : Le profil utilisateur :* Ce troisième chapitre est consacré au profil utilisateur, sa définition, sa modélisation, sa représentation, ses différentes dimensions et un survol sur ses utilisations dans le web 2.0 et notamment ses exploitations dans le Tagging collaboratif.

Deuxième partie : Approche de filtrage de tags à base du profil utilisateur

Nous présentons dans cette partie notre approche de filtrage de tags basée sur le profil utilisateur. Nous consacrons le chapitre IV pour la présentation de notre approche, et le chapitre V pour la mise en œuvre de celle-ci et son évaluation.

- *Chapitre IV : Présentation de l'approche :* Dans ce chapitre, nous présentons nos motivations et le principe général de l'approche, nous détaillons par la suite la partie concernant le modèle utilisateur, puis la partie concernant la pondération des tags.
- *Chapitre V : Tests et évaluations :* pour que notre contribution soit bien appréciée, nous proposons une première implémentation de l'approche proposée. Nous décrivons notre collection de test et la démarche suivie pour l'évaluation. Enfin, nous présentons et discutons nos résultats.

Nous terminons ce mémoire par une conclusion générale, où nous synthétisons le travail effectué, les apports de notre approche quant aux objectifs définis. Nous présentons enfin les perspectives envisageables pour notre travail.

Première Partie
Etat de l'Art

Les informations sur le web sont fortement distribuées, d'une grande hétérogénéité et souvent très peu structurées. Dans ce contexte, il est nécessaire de mettre à disposition des utilisateurs des outils pour comprendre, manipuler et partager ces informations. Les outils d'annotation visent à améliorer la communication et l'interopérabilité sur le Web. Ces annotations permettent d'associer des notes de lectures aux documents et de partager de l'information. Grâce aux techniques d'annotation et du Tagging, le lecteur devient aussi rédacteur. On passe du "one-to-many" (un rédacteur et des millions de lecteurs) au "many-to-many" (tout utilisateur du Web est Lecteur/Rédacteur). Dans cette partie du mémoire, nous aborderons la notion de l'annotation et les concepts qui lui sont liés ainsi que le Tagging collaboratif et le profil utilisateur.

Chapitre I : Les Annotations

Il y a plus de texte écrit sur un visage que dans un volume de la Pléiade et, quand je regarde un visage, j'essaie de tout lire, même les notes en bas de page.
(La lumière du monde, Citations de Christian Bobin)

I.1. Introduction

La pratique d'annotation est très courante, que ce soit sous forme papier ou électronique, nous avons tous l'habitude d'écrire nos commentaires, nos notes pour se rappeler ou expliquer un passage. Elle est utilisée dans différentes fonctions et prend des formes variées, dans l'enseignement, les instituteurs annotent les copies des élèves, en médecine, les médecins commentent les dossiers des patients, etc.

De nombreux outils ont été développés pour annoter les documents numériques, souvent adaptés à des domaines d'application spécifiques et pour des utilisations particulières des documents. **[Bringay, 06].**

Dans ce chapitre, nous mettons le point sur la définition d'une annotation, la structure de l'annotation à savoir l'objet et l'activité annotation, les modèles de représentation de l'objet annotation, les objectifs des annotations selon plusieurs classifications, les catégories de l'objet et de l'activité annotation. Nous citons une classification des outils d'annotations selon les catégories. Puis nous détaillons un type d'annotations à savoir l'annotation sémantique, ses définitions et ses outils.

I.2. Définitions

Dans la littérature, il n'existe pas de définition consensuelle, on trouve plusieurs définitions de l'annotation, différentes mais s'accordent sur le fait que l'annotation est un objet et une activité. Les définitions diffèrent selon les domaines de recherche.

Selon les documentalistes **[Huart, 96],** l'annotation est l'activité du lecteur qui consiste à poser des marques graphiques ou textuelles sur un document papier, et ce suivant plusieurs objectifs.

Dans le domaine des IHMs, l'annotation est définie comme un commentaire sur un objet tel que le commentateur veut qu'il soit perceptiblement distinguable de l'objet

lui-même et le lecteur l'interprète comme perceptiblement distinguable de l'objet lui-même **[Baldonado, 00].**

Pour les psycholinguistiques et cogniticiens, l'annotation est une trace de l'état mental du lecteur et une trace de ses réactions vis-à-vis du document. **[Veron, 97].** Concrétisée par des marques reflétant l'intérêt du lecteur, son activité. On parle de 'Lecture active' **[Adler, 72]** par opposition à une lecture de loisir **[Damas, 02]**.

[Bringay, 03] de sa part, définit l'annotation comme étant une note particulière attachée à une cible. La cible peut être un document, une collection de documents, un segment de document ou une autre annotation. A une annotation correspond un contenu, matérialisé par une inscription, qui est une trace de la représentation mentale que l'annotateur se fait de la cible. Le contenu de l'annotation pourra être interprété à son tour par un autre lecteur. Nous appelons l'ancre ce qui lie l'annotation à la cible (un trait, un passage entouré…).

Quant à **[Atilf, 92]** l'annotation est l'action d'annoter et le résultat de cette action, en définissant l'action d'annoter comme le fait d'accompagner un texte de notes, de remarques, de commentaires. L'annotation indique à la fois une action et un objet.

À travers ces définitions, on peut conclure qu'une annotation est une information graphique ou textuelle attachée à un document et le plus souvent placée dans ce dernier. La place de l'annotation est appelée ancre. Elle peut être aussi une entité distincte du document (ne fait pas partie du document), et n'a de sens que dans le contexte du document. Enfin, elle peut être vue comme étant le fruit d'une lecture active et une tâche cognitive plus intense qu'une simple lecture.

I.3. Structure de l'annotation

L'annotation est à la fois l'action d'annoter et le résultat de cette action qu'est l'objet. Nous parlerons alors d'objet annotation et d'activité annotation.

I.3.1 L'objet Annotation

L'objet annotation est une forme visible ajoutée à un document, une collection de document ou segment de document distinguable du document qu'il annote, dépendant de celui-ci sans le modifier. **[Azouaou, 06].**

I.3.1.1 Composants de l'objet annotation

L'objet annotation est caractérisé par son ancre sur le document et sa forme graphique.

L'ancre : C'est le point d'attachement de l'annotation au document annoté (cible) **[Bringay 03]**. Jaques Virbel dans **[Veron, 97]** en distingue plusieurs niveaux :

- Dans le document
  - Dans la page
    - Dans le texte (souligné)
    - Autour du texte (dans les marges, bas de page…)
    - Par-dessus (texte barré)
    - A côté du texte
  - Entre les pages (intercalaires)
- Hors document (cahier de notes).

A cet effet, il existe deux types d'annotation selon que l'ancre soit à l'intérieur ou à l'extérieur du document : annotations internes et annotations externes.

a. La forme graphique : C'est la forme que peut prendre une annotation sur le document. **[Mille, 05]** fait une liste exhaustive des formes existantes :

- Ajout
  - de texte
  - d'un dessin explicatif
  - d'une marque
    - ✓ unaire
    - ✓ binaire
- Mise en évidence
  - en soulignant
  - en surlignent
  - en entourant
  - en barrant
  - en changeant la couleur

Cependant certaines annotations (non-cognitives) ne possèdent pas de forme visible.

I.3.1.2 Modèles de représentation de l'objet annotation

Les modèles de représentation de l'objet annotation définissent la composition de sa structure et les propriétés qui permettent de l'identifier et de la décrire. Il en existe plusieurs, nous présentons dans ce qui suit, les plus connus.

a. **Le modèle de Matthieu VERON [Veron, 97]:** Veron propose un modèle pour la représentation des annotations composé des propriétés suivantes :

- **La forme :** On parle de l'aspect visuel de l'annotation, de son apparence graphique. Elle dépend des outils mis à disposition de l'utilisateur par le logiciel d'annotation. Exemples : souligné, surligné, marque, etc.
- **Le but de l'annotation:** C'est pourquoi l'annotation est faite. Les buts d'annoter sont différents des buts de la lecture. Le but de l'annotation fixe la sémantique des annotations attachées sur le document.
- **Le lieu de l'annotation**: C'est l'endroit où est placée l'annotation. Il diffère selon le but de cette annotation.
- **L'auteur :** C'est la personne ou l'annotateur qui annote le document. C'est une information pertinente, notamment dans le cas où plusieurs personnes peuvent annoter le même document.
- **L'histoire :** C'est la propriété qui décrit le moment où l'annotation a été placée sur le document. Elle comporte la date de création de l'annotation, mais aussi les dates de modification. La gestion de l'historique des annotations est comparable à une gestion des différentes versions d'un document.
- **Le support de l'annotation :** C'est l'entité concernée par l'annotation. Elle peut être le document en-entier, une phrase, un terme, un mot, une image.

b. **Le modèle de DENOUE [Denoue, 00]:** Dans le contexte de recherche d'information sur le WEB, l'annotation pour **[Denoue, 00]** est structurée en deux parties : l'ancre et le commentaire, l'ancre permet d'attacher l'annotation à une partie précise du document, le commentaire est un ensemble d'attributs facultatifs (auteur, sujet, date de création…).

c. **Le modèle ANNOTEA [Kahan, 02] :** ANNOTEA est un outil collaboratif d'annotations basé sur le web, où les annotations sont considérées comme une classe de métadonnées. Celles-ci sont externes aux documents, et peuvent être stockées dans un ou plusieurs serveurs d'annotation. Dans ce modèle les annotations ont la structure suivante :

- **Type** (***rdf:type***): indique l'objectif de l'auteur en créant l'annotation.
- **Annote** (***annotates***): la relation entre une annotation et la ressource à laquelle s'applique cette annotation, la relation inverse est 'possède une annotation'.
- **Corps** (***body***): le contenu de l'annotation.
- **Contexte** (***context***): indique la partie du document où l'annotation est attachée (le paragraphe, la phrase…)
- **Auteur** (***dc:creator***): le créateur de l'annotation (l'annotateur) ;
- **Crée** (***created***): la date et l'heure à laquelle l'annotation a été créé ;
- **Modifié** (***dc:date***): la date et l'heure de la dernière modification de l'annotation ;
- **Lié** (***related***): la relation entre une annotation et d'autres ressources qui peuvent enrichir cette annotation.

La figure 01 illustre la représentation d'une annotation dans ANNOTEA.

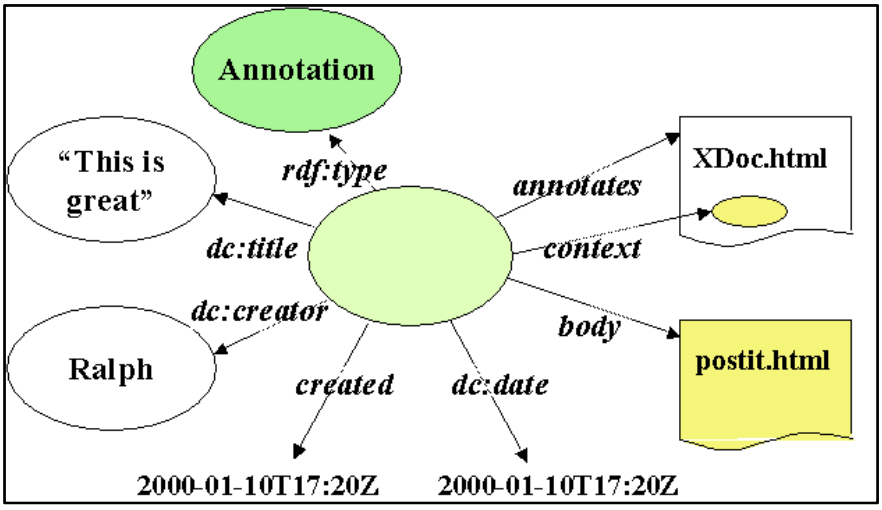

Fig01 : *Représentation d'une annotation dans Annotea*

d. Le modèle du projet MICA : [Desmoulins, 02] propose la structure suivante de l'objet annotation :

- **Ancre** : représente le lieu d'attache de l'annotation, décomposé en URL du document et l'emplacement de l'annotation dans le document.
- **Forme de visualisation** : il s'agit de la forme graphique de l'annotation,
- **Contexte** : les propriétés du contexte sont :
  - **Pourquoi ?** L'objectif de l'annotation ;

- **Pour qui ?** Le destinataire de l'annotation (publique ou à un destinataire particulier) ;
- **Par qui ?** L'auteur de l'annotation ;
- **Quand ?** Date et heure de création de l'annotation ;
- **Où ?** L'annotation a été placée ;
- **Quoi?** Domaine de validité du document, des documents de cours ne sont valables que pour certaines disciplines mais ne le sont pas pour d'autres.

- **Valeur** : c'est le contenu de l'annotation. Ce contenu peut être libre ou prédéfini ou les deux. Il peut aussi faire référence à des notions de cours ou à d'autres documents.
- **Accès** : il s'agit des privilèges d'accès pour la consultation de l'annotation, on trouve trois types :
  - Privé : réservé pour l'auteur de l'annotation ;
  - Groupe : l'accès est réservé pour un groupe prédéfini d'utilisateurs ;
  - Public : l'accès à l'annotation est libre pour tous les utilisateurs.

Les modèles cités ci-dessus ont en commun certaines propriétés fondamentales à savoir l'auteur, la date de création et l'objectif, mais comportent des différences de structure qui dépendent de l'objectif pour lequel l'annotation a été crée.

I.3.2. L'activité Annotation

L'activité annotation est le processus de création de l'objet qui vise la réalisation d'un objectif. Ce processus peut être manuel, semi-automatique ou automatique. Ceci sera détaillé dans la section catégories de l'activité annotation.

[Huart, 96] formalise l'activité annotation en donnant un schéma de création pour chaque forme d'annotation, le tableau ci-dessous montre quelques exemples :

Forme d'annotation	Schéma de création
Mise en évidence en surlignant	1- Choix du passage à mettre en valeur 2- Choix de la mise en valeur
Ajout d'une marque binaire (lien)	1- Choix de l'ancre 2- Choix du passage ou du point en relation
Ajout d'une marque	1- Choix de l'ancre 2- Ajout (éventuel) d'un commentaire
Ajout d'un dessin explicatif	1- Choix de l'ancre 2- Ajout du dessin

Tableau01 : *Exemples de schéma de création d'annotations* **[Azouaou, 06]**

Certains auteurs tels que **[Marshall, 97]** et **[Huart, 96]** se sont intéressés à l'étude des correspondances entre les formes graphiques et les sémantiques d'annotation voulues par l'annotateur. Les auteurs ont identifié l'existence d'habitudes d'annotation qui sont communes et partagées dans un groupe d'annotateurs. Elles représentent la correspondance entre les formes utilisées pour annoter et l'objectif visé. Par exemple, souligner ou mettre en surbrillance est une procédure pour signaler une attention future, et une notation appropriée sur la marge ou prés des figures ou d'équations a comme fonction la résolution de problèmes (problem working).

I.4. Sémantique de l'annotation (objectifs)

L'objectif de l'annotation est un point important du moment qu'il justifie la présence de celle-ci sur le document. Le lecteur laisse des traces dans les annotations, selon son objectif de lecture, qui sont significatives de sa lecture des documents **[Bringay 04]**. Plusieurs auteurs ont identifié les objectifs d'une annotation:

I.4.1. Les objectifs selon C. Marshall [Marshall, 98]

Une étude sur les annotations créées par des étudiants dans leurs supports de cours universitaires a été effectuée par Marshall. Dans son étude, Marshall a observé les caractéristiques des annotations et en a déduit les objectifs suivant :

- **Procédure pour signaler une attention future** : l'étudiant annote pour se souvenir qu'il devrait relire le passage annoté ou faire ultérieurement une tâche particulière.

- **Aide Mémoire et indication d'emplacement** : souvent en surlignant (ou même soulignant) un passage du texte, signifie que celui-ci est important et qu'il doit être mémorisé.
- **Résolution du problème (Problem-working)** : par risque d'oubli ou de perte d'enchainement d'idées, un étudiant préférera noter sa solution à un problème sur les marges.
- **Interprétation** : les annotations interprétatives peuvent être des traductions des mots de langues étrangères, des synonymes des mots difficiles ou des explications du sens d'un texte donné.
- **Trace visible de l'attention du lecteur** : généralement un texte difficile (philosophique par exemple), peut être l'objet d'un volume important d'annotations pour mieux gérer la charge cognitive de la sémantique du texte.
- **Réflexion fortuite des circonstances matérielles de la lecture** : c'est une annotation indépendante du texte en question, Marshall a cité l'exemple d'une opération arithmétique trouvée dans un document de philosophie.

I.4.2. Les objectifs selon J. Virbel [Veron, 97]

Virbel a définit les objectifs des annotations dans le cadre du projet de réalisation d'un poste de lecture active à la Bibliothèque Nationale de France comme suit:

- **Classifier :**
  - **Hiérarchiser** : C'est l'affectation d'une valeur numérique à un objet afin de le situer sur une échelle relativement à d'autres objets.
  - **Architecturer** : Mettre en évidence la structure logique du document (chapitre, section…).
  - **Contextualiser :** Créer un passage hors duquel l'annotation peut ne plus avoir le sens voulu, (exemple : Souligner un mot, et puis donner le passage).
- **Compléter :**
  - **Reformuler** : Donner une nouvelle représentation de l'objet support (exemple donner le terme entier d'une abréviation).
  - **Commenter** : Faire un commentaire sur l'objet support, (exemple : critique, explication).
  - **Documenter** : Ajouter à l'objet support, un autre objet qui complète celui-ci (exemple : photo pour mieux expliquer un passage).

- **Corréler :** Il s'agit de lier deux parties entre elles.
- **Planifier :** Programmer une action : à traduire, à relire, à analyser…

I.4.3. Les objectifs selon [Mille, 05]

Un travail plus récent de Mille a laissé naitre un ensemble d'objectifs. Il s'est appuyé sur ses propres expériences et les classifications précédentes**.**

- **Restructurer :** Donner un titre, Hiérarchiser, Synthétiser, Reformuler
- **Ajouter une remarque personnelle :**
  - Critiquer (Positivement ou Négativement)
  - Exprimer une idée connexe : Développer, Compléter, Ajouter un exemple, Résoudre un problème, Expliquer textuellement/graphiquement
  - Faire référence à un autre document
- **Catégoriser**
  - Objectivement : Par type prédéfini, Par similarité de forme
  - Subjectivement : Par valeur d'importance (du passage, pas de l'annotation), Par similarité de sens.
- **Créer une relation entre deux passages**
- **Planifier une action :** Approfondir, Réviser (Supprimer, Insérer, Reformuler ou Déplacer un passage)
- **Soutenir l'attention**

De ces différentes visions de l'objectif des annotations, nous retenons que celui-ci est primordial pour retrouver la sémantique voulue par l'annotateur. Les auteurs ont mené leurs études dans des contextes différents et des annotateurs différents (étudiants, apprenant, lecteur en bibliothèque…). La sémantique de l'annotation dépend donc de la forme et aussi du contexte.

I.5. Catégorisation des annotations

Il existe dans la littérature des catégories d'annotation liées à l'objet et d'autres liées à l'activité annotation.

I.5.1. Les catégories de l'objet annotation

Il existe trois catégories de l'objet annotation, ces catégories sont valables pour l'objet annotation informatique, on parlera de destinataire humain ou logiciel **[Azouaou, 06]**.

I.5.1.1. L'annotation cognitive

Si l'annotation est destinée à être lue et interprétée par l'agent humain, celle-ci est appelée « annotation cognitive », elle doit avoir une forme visible, perceptible et distinguable du document qui la porte. Cette catégorie d'annotation nécessite un effort cognitif et intellectuel de la part de l'agent humain. Dans le cas contraire, lorsque l'annotation n'est pas destinée à un agent humain, l'annotation est invisible et appelée « annotation non cognitive ».

I.5.1.2. L'annotation computationnelle

Lorsque l'annotation est destinée à être interprétée par un agent logiciel, elle est appelée « annotation computationnelle ». Elle est aussi souvent appelée métadonnée. L'annotation computationnelle vise à décrire des ressources informatiques.

I.5.1.3. L'annotation sémantique

Ce type d'annotation sert à décrire des ressources en rajoutant une couche de connaissances liée à ces ressources. Cette couche de connaissances peut être une ontologie ou un réseau sémantique de concepts, l'annotation sémantique est liée à un concept donné appartenant à la représentation de connaissances. Nous détaillons ce type d'annotation dans la section I.7.

I.5.2. Les catégories de l'activité annotation

Tout processus de l'activité annotation passe par trois étapes (sous-processus) :

- Choix de l'ancre et de la forme de l'annotation ;
- Spécification des propriétés de l'annotation ;
- Choix de la cible dans l'ensemble de la représentation formelle ou non.

A partir de ces trois sous-processus, il a été défini trois catégories de l'activité annotation:

I.5.2.1.L'annotation manuelle

Dans ce cas, les trois sous-processus sont exécutés manuellement par l'annotateur, tout le processus est donc à la charge de celui-ci. Pour le cas d'une annotation sémantique,

le processus manuel devient pesant du fait que l'annotateur doit spécifier pour chaque annotation le concept qu'elle représente dans la couche de connaissances.

I.5.2.2. L'annotation semi-automatique

Dans ce type d'annotation, l'un des trois sous-processus est exécuté par un outil d'annotation. Elle est utilisée particulièrement dans le web sémantique. A cet effet, différents outils sont proposés : Melita **[Ciravegna, 02]**, Ontomat **[Handschuh, 03]**.

I.5.2.3. L'annotation automatique

Ce type d'annotation signifie que les trois sous-processus sont exécutés automatiquement par un outil d'annotation. Celui-ci sélectionne lui-même les ancres, crée les annotations, les enregistre et les affiche (dans le cas des annotations cognitives). L'un des outils les plus connus, est la barre de recherche GoogleToolbar qui permet de surligner avec différentes couleurs les mots clés tapés par l'utilisateur.

I.6. Les outils d'annotations

Il existe plusieurs outils d'annotation cognitive comme IMarkup. D'autres outils d'annotation computationnelle comme **[Kalyanpur, 04]** qui représente l'annotation avec un langage formel interprétable par les agents logiciels. Melita et Ontomat pour l'annotation semi-automatique.

[Azouaou , 05] classifie les outils d'annotation selon les types d'objets et d'activités annotation crées. Cette classification est illustrée dans le tableau suivant. Nous abordons d'autres classifications des outils d'annotation dans l'Annexe A.

Type de l'objet annotation		**Type de l'activité annotation**		
Sémantique	**Cognitive Computationnelle**	**Manuelle**	**Semi-automatique**	**Automatique**
Annotation non sémantique	Cognitive et non computationnelle	Imarkup [iMarkup-Solutions-Inc 2004], Web-notes [Ronchetti et al.2002], CoNote [Davis et al.1995], WebAnn [Brush et al.2002], epost [Brust et al.2002].		Barre de recherche Google
	Non cognitive et computationnelle	Index manuel dans les bibliothèques	MyAlbum [Wenyin et al.2001], Annotate [Plaehn et al.2000]	Moteur de recherche de Google
	Cognitive et computationnelle	Knowledge Pump [Glance et al.1999], Xlibris [Golovchinsky et al.1999]		Pages cachées de Google [Google corporate 2004]
Annotation sémantique	Cognitive et non computationnelle	Annotea et Amaya [Ciravegna et al.2002], Yawas[Denoue et al.2000], ThirdVoice[1999], Mark-Up [McMahon et al.2003]		
	Non cognitive et computationnelle	Edutella [Nedjl et al.2002], OntOmat [Handschuh et al.2002], SHOE [Heflin et al.2000], HTML-A [Decker et al.1999], WebKB [Martin et al.1999], Karina [Crampes et al.2000]		AeroDAML [Kogut et al.2001]
	Cognitive et computationnelle	Mangrove [McDowell et al.2003], SMORE [Kalyanpur et al.2004]	MnM [Vargas-Vera et al.2002], Melita [Ciravegna et al.2002], Teknowledge [Tallis 2003], IMAT [De Hoog 2002]	KIM [Popov et al.2003], MnM [Vargas-Vera et al.2002], Magpie [Gaasterland et al.2000], COHSE [Goble et al.2001]

Tableau02 : *Classification des Outils d'annotations existants,* ***[Azouaou, 05]***

Nous remarquons que si le processus d'annotation nécessite un effort considérable de la part de l'utilisateur, il est essentiel d'automatiser partiellement ou complètement l'activité annotation.

[Bringay, 06] a également classifié les outils d'annotation selon le type d'objet à annoter, en cinq catégories :

- Outils d'annotation de page Web : Annotea et IMarkup,
- Outils d'annotation de documents pouvant être utilisés pour des lectures personnelles ou pour la co-construction de documents : Adobe Acrobat et Microsoft Office (Word),
- Outils d'annotation de contenus multimédias : Debora permet l'annotation de livres scannés,
- Outils d'annotation de dispositifs mobiles, comme les PDA, les e-books, les tablettes PC : Adobe eBook Reader, Xlibris,
- Autres applications : Connotea permet d'organiser un ensemble de ressources disponibles en ligne par la création de signets (marque-pages pouvant être vues comme des annotations), Kinoa permet d'annoter une bibliothèque de documents, Magpie permet, en fonction d'une ontologie, de surligner des éléments dans des ressources pédagogiques.

I.7. L'annotation sémantique

L'objectif du web sémantique est de rendre les ressources du web interprétables par la machine, accessibles et utilisables par des programmes et des agents logiciels. En effet les documents traités par le web sémantique contiennent des informations formalisées pour être traitées automatiquement et non pas interprétées uniquement par les humains.

I.7.1. Définitions

L'annotation sémantique sert à décrire des ressources du web en ajoutant une couche de connaissances. D'après **[Demontils, 02]**, les annotations sémantiques sont le plus souvent attachées au document et ne possèdent pas d'ancrage particulier. Elles sont destinées à être traitées par des machines, leur objectif majeur est de désambiguïser le document pour un traitement automatique. **[Amardeilh, 07]** définit l'annotation sémantique comme étant une représentation formelle d'un contenu, exprimée à l'aide de

concepts, relations et instances décrits dans une ontologie, et reliée à la ressource documentaire source.

I.7.2. Les langages d'annotation sémantique

L'expression des annotations se fait à l'aide de plusieurs langages, RDF (Ressource Description Framwork), Topic Maps, RDF-Schéma et OWL. D'autres langages appelés précurseurs tels que HTML-A (HyperText Markup Language- Annotation) (Fig 02), SHOE (Simple HTML Ontology Extension) sont nés, ils représentent des extensions de HTML et permettant l'insertion des annotations sémantiques pour décrire des ressources Web. OWL est utilisé pour décrire les ontologies dont les concepts sont utilisés pour l'annotation et RDF-S pour décrire des ontologies légères (utilisées dans le Tagging). (Certains langages d'annotation sémantique sont détaillés dans l'Annexe A).

```
<html>
<head><Title>Le Clan coppola</Title>
<A ONTO="Personnalité:FFCoppola"/>
</head>
<body>
Francis Coppola naît le <A ONTO="Personnalité[dateNaissance=body]">7 avril 1939</A> à <A
ONTO="Personnalité[lieuNaissance=body]">Detroit</A>, dans le <A
ONTO="Personnalité[lieuNaissance=body]">Michigan</A>.
</body>
</html>
```

Fig 02 : *Exemple d'une annotation sémantique avec HTML-A*

Nous situons l'annotation du web sémantique par rapport aux catégories définies précédemment comme suit :

- elle est destinée à des agents logiciels, elle est donc computationnelle ;
- elle ne possède pas de forme graphique visible à l'humain, elle n'est donc pas cognitive ;
- elle possède une sémantique formelle grâce à/aux ontologie(s) qui la structure.

I.8. Conclusion

Tout au long de ce chapitre, nous avons parcouru les différentes notions liées aux annotations : leurs objectifs et leur catégorisation ainsi que les outils d'annotation existants. Nous avons abordé par la suite, l'annotation sémantique et les langages utilisés pour sa réalisation.

Dans un environnement tel que le web, les annotations sont souvent partagées entre plusieurs utilisateurs, et par fois même, créées en collaboration entre plusieurs utilisateurs.

On parle ainsi des annotations sociales ou plus encore du Tagging collaboratif qui est le concept associé à ce type d'annotation. Nous abordons plus en détail, le Tagging collaboratif dans le chapitre suivant.

Chapitre II : Le Tagging Collaboratif

Il n'y a point de véritable volonté sans liberté
(Jean-Jacques Rousseau, Ém. IV)

II.1. Introduction

Marquage collaboratif, étiquetage collaboratif, Tagging social, annotations sociales ou Tagging collaboratif, différentes appellations désignant toutes ce phénomène qui est apparu ces dernières années et qui ne cesse de gagner une popularité sur le web. Marquer un contenu par des termes descriptifs est une manière d'organiser ce contenu pour une navigation future, un filtrage ou une recherche. Le Tagging collaboratif est devenu un moyen de plus en plus courant pour le partage et l'organisation du contenu web. D'autres concepts découlent du Tagging telles que la Folksonomie et l'indexation collaborative.

Actuellement, de nombreux sites offrent la possibilité de tagguer du contenu. Divisés en catégories, il y a ceux spécialisés dans le partage des papiers scientifiques ou de références bibliographique tels que Connotea, d'annotation de photos comme Flickr ou de vidéos comme Youtube et Dailymotion en encore de signets (bookmarks) tel que Delicious.

Dans ce chapitre, nous allons définir le Tagging collaboratif, les Folksonomies et les différentes structures de l'action du Tagging. Nous parlerons également des systèmes du Tagging collaboratif et leurs principales propriétés. Enfin, nous mettrons l'accent sur les travaux de recherche actuels dans ce domaine, notamment dans la recherche d'information (RI) et les travaux sur le rapprochement d'ontologies et de folksonomies.

II.2. Définitions

II.2.1. Le Tagging collaboratif

On désigne par *Tagging Collaboratif* le processus qui consiste à associer un ou plusieurs "tags" (mot clé) à un document numérique (page web, photo, vidéo, billet de blog) dans un environnement multi utilisateurs.

Selon **[Golder, 06]** *le Tagging Collaboratif* décrit le processus par lequel plusieurs utilisateurs ajoutent des métadonnées à un contenu partagé.

II.2.2. Le Tag

Un *Tag* ou étiquette est un mot clé librement choisi par un utilisateur pour décrire un objet dans le web (document texte, image, fragment d'un document).

Un tag peut être vu simplement comme étant un jeu de mots-clés librement choisi par les utilisateurs. Et le fait que ces derniers ne soient pas spécialistes de l'information, leurs tags ne suivent aucune indication formelle. Cela signifie que ces mots peuvent être catégorisés avec n'importe quel mot définissant une relation entre la ressource et un concept issu de l'esprit de l'utilisateur. Un nombre infini de mots peut être choisi, dont quelques-uns sont issus de représentations évidentes tandis que d'autres ont peu de signification en dehors du contexte de l'auteur du tag, **[Guy, 06].**

En se basant sur la communauté du Tagging dans Delicious, **[Golder, 05]** spécifie sept (07) fonctions que peut avoir un tag :

1- Identifier de quoi l'objet s'agit-il, son thème ou sujet : vacances, hiver ;
2- En plus d'identifier le thème de l'objet (ou du contenu), un tag peut identifier ce qu'est l'objet lui-même : une photo, un blog… ;
3- Identifier à qui l'objet appartient ;
4- Description ou détail de tags existants: Certains tags n'ont aucune signification seuls. Ils n'ont de sens que quand ils sont associés à d'autres tags. Leur fonction est donc d'apporter plus de détails ou de description à des tags existants. Tel est souvent le cas avec les nombres comme par exemple le 10 de l'expression top 10.
5- Identifier des qualités ou des caractéristiques du contenu : c'est le fait de dire que tel contenu est 'comique', 'funny' ou 'horrible'… ;
6- Auto référence du tag : dans ce cas, le tag illustre une relation entre l'utilisateur et le contenu. C'est le cas des tags commençant par 'mon' ou 'my' : myphoto, mon_enfant… ;
7- Aide-mémoire : il s'agit de planifier une tâche donnée, 'à lire', 'à revoir'…etc.

II.2. 3. La Folksonomie

II.2.3.1. Définition

La *Folksonomie* est un terme anglais introduit en 2004 par Vander Wall, exprimant l'idée d'une classification (Taxonomie) faite par les utilisateurs (Folks). Une folksonomie est le résultat de la collecte de données du Tagging pour un groupe donné, elle est donc liée à un site communautaire bien particulier : par exemple, la folksonomie de Flickr est différente de celle de

Youtube. Cependant elle est souvent confondue à son processus de création qu'est *le Tagging collaboratif*.

Les folksonomies sont des séries de métadonnées créées en collectif par les utilisateurs pour catégoriser et retrouver les ressources en ligne **[Broudoux, 06].** Une folksonomie est différente d'une taxonomie car, d'une part, elle n'est pas contrainte par des relations hiérarchiques, et d'autre part, elle n'est pas conçue par des experts. Il ne s'agit pas non plus d'une ontologie. Une ontologie est un ensemble structuré de concepts, alors qu'une folksonomie ne possède qu'une structure émergente, floue, et non contraignante (exemple un utilisateur peut utiliser un tag dans un sens totalement différent des autres utilisateurs).

II.2.3.2. Avantages et inconvénients des folksonomies

La liberté du choix des tags offre aux folksonomies un certain nombre d'avantages, cette même caractéristique est à double tranchant et provoque des limites : **[Mathes, 04],** résume ce paradoxe en une autre phrase paradoxale : « *Une folksonomie représente en même temps ce qu'il y a de meilleur et de pire dans l'organisation de l'information* ». Un vocabulaire non contrôlé, ne peut qu'avoir des limites :

- **Ambiguïté :** on parle d'ambiguïté quand un même tag dénote deux concepts différents : le terme orange pour le fruit, la couleur ou la société française de télécommunication.
- **Hétérogénéité :** c'est lorsqu'un un tag se présente en différentes formes, Hétérogénéité englobe la variation d'écriture : New York et New_York, les synonymes : 'mac' et 'macintosh', ou télévision et TV, l'utilisation ou non du pluriel : 'flower' et 'flowers', et le multilinguisme : 'chat' et 'cat'.
- **Info-pollution ou Spamming :** certains utilisateurs malveillants, peuvent nuire en inondant les contenus de tags inadéquats. Effet indésirable notamment pour les sites du e-commerce.

Or les folksonomies marchent bien, certes elles ne présentent pas que des inconvénients ; en quoi donc consiste leurs force ?

- **Peu coûteuse :** Puisque la folksonomie est réalisée par les utilisateurs finaux et non pas par des professionnels.
- **Tagging continu, folksonomie dynamique** : Elle est mise à jour automatiquement et en permanence (au fur et à mesure de l'activité des utilisateurs).
- **Folksonomie intuitive** : puisque elle est le fruit de collaboration de simples utilisateurs.

- **Améliore les résultats de recherche** : Avec une folksonomie on peut tomber sur des documents qu'un moteur de recherche classique aurait pu ignorer (documents non-indexés par le moteur).
- **Utilisée pour la veille :** Un point fort des folksonomies et la possibilité de leurs utilisations pour la traque (tracking) : Technorati utilise le tracking de termes précis, on retrouve les blogs où sont employés ces termes.

II.2.3.3. Folksonomie et indexation collaborative

Cette classification faite par les utilisateurs est souvent appelée *Indexation collaborative* des documents. Peut-on vraiment parler d'indexation sachant que l'indexation classique se fait par des experts, et elle est d'une grande objectivité. Cependant l'indexation collaborative, née du processus du Tagging collaboratif, n'est pas forcement objective et peut présenter d'autres inconvénients. **[Le Deuff, 06]** a édicté des règles d'une bonne indexation en se basant sur les travaux du consultant en science de communication Ulises MEJIAS :

- L'utilisateur doit penser collectivement : les tags sont certes personnels mais peuvent également être utilisés par d'autres ;
- Employer le pluriel pour définir des catégories. Le pluriel est plus approprié car la catégorie peut contenir différentes variations ;
- Ne pas employer de majuscules, à moins que le mot ne puisse être compris sans ;
- Utiliser l'*underscore* pour définir un groupe de mots ;
- Inclure des synonymes afin d'éviter les confusions ;
- Observer et utiliser les conventions d'indexation des sites et des réseaux sociaux utilisés.

II.3. Structure d'une action du Tagging collaboratif

La principale structure d'une action de Tagging est la structure tripartite, cependant il existe d'autres types de structure à savoir la structure tripartite avec liens et la structure quadripartite.

II.3.1. Structure tripartite de base

Toute action du Tagging est composée de trois principales entités : l'utilisateur, la ressource et le/les tags utilisés, (Fig 03), **[Marlow, 06].**

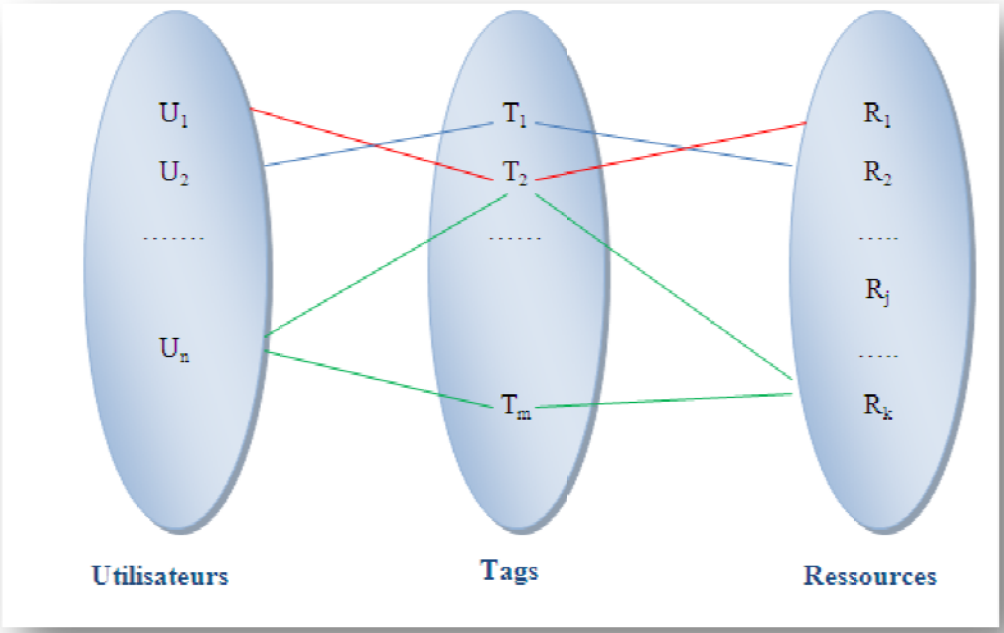

Fig 03 : *Structure tripartite d'un ensemble d'actions du Tagging collaboratif,* ***[Halpin, 07]***

L'utilisateur : C'est celui qui taggue, il est un inscrit dans le site communautaire, donc connu par son profil ;

Le tag : Le ou les tags, est l'ensemble de mot clés utilisés pour décrire la ressource ;

La ressource : C'est un contenu partagé sur le web, il peut être un document texte, photo, vidéos…

Une instance de cette relation est composée d'un utilisateur, un ou plusieurs tags et une ressource : Tagging (utilisateur, ressource, tag).

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

Dans un modèle de Tagging, des liens peuvent être présents entre les ressources (tels que les liens entre pages web), mais aussi entre les utilisateurs (réseau social). Nous pouvons voir ceci dans le modèle conceptuel de **[Marlow, 06],** où les tags sont représentés sous forme d'arêtes reliant les ressources aux utilisateurs, les liens entre utilisateurs ou entre ressources sont représentés en pointillés. (Fig 04).

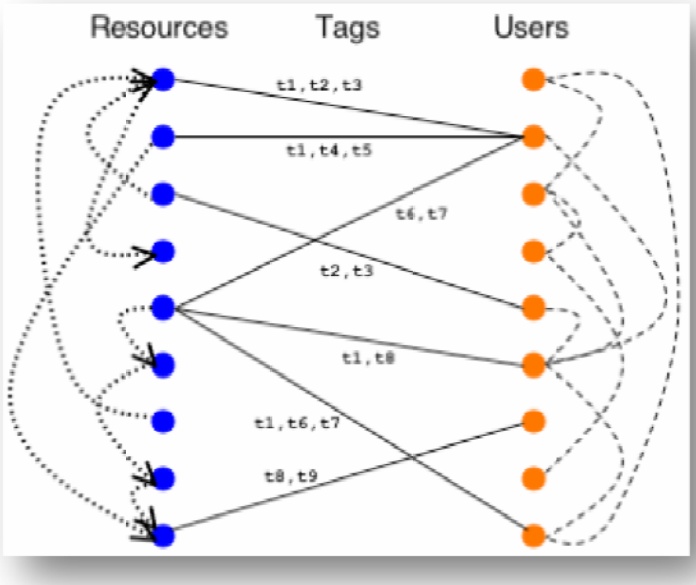

Fig 04 : *Modèle conceptuel d'un système de Tagging* ***[Marlow, 06]***

Même si ces liens n'existent pas explicitement, on peut dire qu'il existe toujours des relations implicites entre ressources à travers les utilisateurs les tagguant (utilisateurs communs), et de même pour les utilisateurs à travers les ressources qu'ils tagguent (ressources communes entre un ensemble d'utilisateurs). On peut parler dans ce contexte d'émergence de relations.

II.3.3. Structure quadripartite

Une extension de la structure tripartite a été proposée dans **[Gruber, 07]** qui ajoute la notion de *source* faisant référence à la signification du tag (Fig 05). La structure d'une action du Tagging est dans ce cas un quadruplet, le triplet (utilisateur, ressource, tag) est enrichi de l'entité *source*. Le modèle de Gruber (utilisateur, ressource, tag, source) est utilisé dans le Tagging guidé par ontologie (Tagging sémantique).

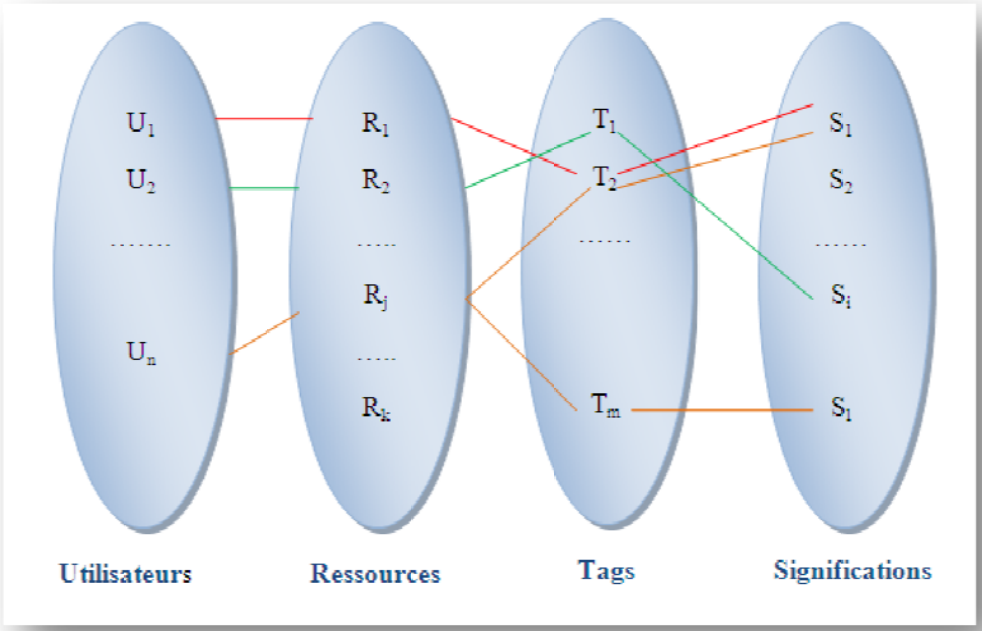

Fig 05 : *Structure quadripartite d'un ensemble d'actions du Tagging collaboratif*

II.4. Propriétés d'un système du Tagging collaboratif

Un système du tagging collaboratif permet à ses utilisateurs d'attribuer des mots clés à des contenus partagés sur le web, nous citons ci-dessous ce qu'offre un tel système pour l'utilisateur :

- Liberté du choix des mots clés (tags), en effet aucun contrôle n'est mis en œuvre pour guider ou forcer l'utilisateur excepté le recours aux ontologies et dans certains cas la suggestion de tags estimés adéquats par le système ;
- Tagguer ses propres ressources, mais aussi les ressources créées et tagguées par d'autres utilisateurs ;
- L'utilisateur doit être identifié, donc inscrit au préalable, les informations requises changent d'un système à un autre. D'une manière générale, l'utilisateur est amené à introduire son nom, prénom, nom d'utilisateur, sexe et une petite description libre de l'utilisateur ;
- L'utilisateur peut associer plusieurs tags pour la même ressource ;
- Un même tag peut être associé à différentes ressources par différents utilisateurs ;
- Une même ressource est tagguée par plusieurs utilisateurs ce qui donne l'aspect collaboratif.

II.5. Tagging vs Annotation

Le Tagging peut être comparé à l'activité annotation qui peut être manuelle ou semi-automatique du moment qu'il nécessite la contribution de l'utilisateur. C'est aussi un acte volontaire et libre où l'utilisateur décrit les ressources d'une manière généralement subjective, chose qui n'est pas toujours évidente pour les annotations notamment l'automatiques où les annotations sont réalisées par l'outil. Le Tagging n'a pas d'ancre, il concerne la ressource entière. L'aspect collaboratif est apparent pour le Tagging (sites communautaires ou réseaux sociaux), quant à l'annotation, un utilisateur donné peut utiliser son propre outil pour annoter, sans pour autant faire participer d'autres utilisateurs. Nous estimons donc que le Tagging est un type d'annotation, volontaire et subjectif, où seront sauvegardées trois entités : l'utilisateur, la ressource et les tags.

Ci-dessous, un tableau résumant les principales caractéristiques du Tagging et de l'annotation :

Caractéristique	Tagging	Annotation
Automatique	Non	Oui
Semi-automatique	Oui	Oui
Manuelle	Oui	Oui
Possède une ancre	Non	Oui
Liberté des choix	Oui	Oui sauf pour l'automatique
Aspect collaboratif	Oui	Pas toujours
Sauvegarde de l'utilisateur	Oui	Pas nécessaire

Tableau 03 : *Comparaison entre Tagging et Annotation*

II.6. Etude des systèmes du Tagging collaboratif

Le Tagging collaboratif est un phénomène récent sur le web, il a gagné beaucoup de popularité et suscité l'intérêt de plusieurs chercheurs. Malgré qu'il soit l'objet de nombreux travaux, il demeure peu étudié et peu compris.

Il existe une multitude de systèmes mettant en œuvre le Tagging Collaboratif, nous citons des exemples dans l'Annexe B.

Dans cette section nous essayons de classifier les différentes études menées sur les systèmes du Tagging collaboratif.

II.6.1. Etude de la dynamique des systèmes du Tagging

Les systèmes du Tagging social sont devenus de plus en plus populaires au cours des dernières années, les pratiques des utilisateurs ont été peu étudiées et peu comprises jusqu'à présent. Cependant, la compréhension du comportement peut contribuer à une compréhension approfondie du phénomène du Tagging. L'étude de la dynamique de ces systèmes est l'une des voies explorées par les auteurs. La dynamique de tels systèmes peut se mesurer en un ensemble de facteurs entre autres : les activités des utilisateurs (fréquences d'utilisation du système), variations du nombre de tags utilisé par utilisateur ou par ressource, les types de tags…etc.

Une première étude de la dynamique a été menée dans **[Golder, 05]**, où les auteurs ont observé que les utilisateurs du système étudié (Del-icio-us), présentent une grande variété dans leurs jeux de tags, les uns ont de nombreux tags, les autres peu. Les tags eux-mêmes varient du point de vue fréquence d'utilisation. Cette étude expérimentale a également démontré qu'après un certain seuil, la fréquence de chaque tag est fixée par rapport à l'ensemble de tags utilisés.

[Halpin, 07] a également fait une étude de la dynamique des systèmes du Tagging, il a conclut que les tags les plus utilisés pour annoter une ressource restent les mêmes après une certaine durée.

Un nombre de chercheurs se sont intéressés à l'étude des facteurs influençant les tags résultants, le comportement des utilisateurs, leurs motivations quant au choix des tags selon différents types de ressources.

Markus Heckner dans **[Heckner, 08]** fait une étude comparative de quatre sites populaires du Tagging social, à savoir Delicious pour les signets, Connotea pour les articles scientifiques, Flickr pour les photos et Youtube pour les vidéos, en se posant les questions :

- Est-ce que la fonction du Tagging diffère d'un type de ressource à un autre ?
- Quelle est la relation entre le titre de la ressource et le tag choisi ?
- Les tags utilisés dans ces différentes plateformes sont t-ils subjectifs (aspect personnel) ou objectifs ?

De son côté, Cameron Marlow dans **[Marlow, 06]** a donné un ensemble de facteurs pouvant influencer la folksonomie résultante et la dynamique du système. Elle a divisé ces facteurs en deux catégories : Conception de ces systèmes et Motivations des utilisateurs.

La conception des systèmes comporte les facteurs suivants :

- ***Les droits du Tagging :*** le système est soit en *self-tagging* où les utilisateurs ne tagguent que leurs propres ressources, ou en *free_for_all-tagging* où les utilisateurs ont le droit de tagguer d'autres ressources.

- ***Type du Tagging :*** la dynamique d'un système du Tagging peut être influencée selon que le Tagging soit *blind-tagging* (aveugle) où l'utilisateur taggant une ressource ne voit pas les tags attribués par les autres pour cette même ressource. Ou *viewable-tagging* (contraire au blind) ou enfin *suggestive-tagging* où une liste de tags sera suggérée par le système.
- ***Type d'objet à tagguer :*** le type d'objet est un facteur décisif pour le choix des tags, en effet, le Tagging des ressources textuelles peut différer du Tagging des ressources vidéo, audio ou images.
- ***Liens entre ressources :*** c'est soit des liens directs, groupés ou pas de liens. La présence de liens peut conduire à une convergence de tags (tags similaires).
- ***Liens entre utilisateurs (sociaux) :*** même chose que pour les liens entre ressources, la présence des liens entre utilisateurs conduit à une adoption d'une folksonomie localisée.

Les motivations de l'utilisateur affectent aussi les types de tags associés à une ressource donnée, ces motivations peuvent être :

- ***Utilisation future:*** un utilisateur taggue une ressource pour se rappeler d'elle pour une utilisation ultérieure.
- ***Contribution et partage :*** l'utilisateur veut contribuer au Tagging d'une ressource au profit d'une ou de plusieurs autres personnes.
- ***Attirer l'attention :*** l'attention de l'utilisateur est attirée lorsque le système se présente dans un style visible et attirant (exemple les nuages de tags).
- ***Jeux et compétition :*** c'est le cas des jeux sur le Tagging tel qu'ESP Game.

II.6.2. Travaux sur la proposition de modèles et d'algorithmes de suggestion de tags et d'utilisateurs

Ces dernières années le web social et le Tagging sont l'objet d'une recherche active. Une panoplie d'algorithmes de suggestion de tags a été développée. Le but est d'orienter le choix de l'utilisateur pour une meilleure sélection de tags adéquats pour un contenu donné, également suggérer à un utilisateur une liste d'autres utilisateurs partageant les mêmes intérêts que lui et finalement recommander une liste de sites (URLs) pour un utilisateur en se basant sur ses informations personnelles (profil par exemple).

Nous citons en l'occurrence **x.qui.site**, étudié dans **[Amer-Yahia, 08]**, est un système qui capture les comportements des utilisateurs et recommande pour ceux-ci des URLs et d'autres utilisateurs selon les comportements étudiés. **x.qui.site** est utilisé par Delicious.

La suggestion de tags est une forme de Tagging semi-automatique, qui toutefois laisse le choix à l'utilisateur de choisir dans cette liste ou d'ignorer les tags suggérés en introduisant un nouveau tag manuellement.

II.6.3. Découverte de communauté

Le but est de regrouper les utilisateurs en communautés d'intérêts, pour ce faire, les tags sont regroupés en catégories en se basant sur l'analyse en composantes principales (ACP).

Dans **[Abrouk, 10]**, à chaque utilisateur est associé un vecteur $\mathbf{X_i}$= (x_{i1},x_{i2},…..x_{im}) des degrés d'appartenance de l'utilisateur u_i à chaque tag. Ce vecteur représente le positionnement de l'utilisateur dans l'espace des tags. Aussi à chaque tag est associé un vecteur $\mathbf{V_j}$ = (x_{1j},x_{2j},…..x_{nj}) de degrés d'appartenance du tag à chaque utilisateur. Le recours à la technique d'ACP est pour diminuer le nombre de dimensions et ne représenter que les plus importantes (significatives et expliquant mieux la variance des données). Des degrés de corrélation de tags à une composante donnée sont calculés, et ne sont retenus que les tags ayant leurs degrés de corrélation supérieurs en valeur absolue à un seuil prédéfini. Les tags retenus par rapport à une composante, forment une communauté.

Pour déduire les communautés d'utilisateurs, à chacun de ces utilisateurs est associé un vecteur de degrés d'appartenance de celui-ci à chaque communauté de tag. La communauté ayant le degré maximum est celle à laquelle cet utilisateur appartient.

II.7. Tagging collaboratif et recherche d'information

L'un des axes qui semble prometteur et intéresse un grand nombre de chercheurs est celui d'exploiter au mieux les tags dans l'objectif d'améliorer la recherche (sélection) des contenus tagués. Nous citons quelques travaux visant à intégrer les tags pour améliorer la recherche sur le web.

Dans **[Bischoff, 08]**, les auteurs se basent sur l'étude des tags et le comportement de l'utilisateur vis-à-vis du Tagging et la recherche. En d'autres termes, classer les tags en catégories (topic, time, location…), étudier la fréquence d'utilisation de chaque catégorie selon le type de ressources (image, musique, page web…), ces tags existent –ils déjà dans les contenus qu'ils annotent ou apportent-ils une valeur ajoutée ? Quant à l'étude du comportement de l'utilisateur, cela consiste à suivre ses choix lors du Tagging et lors de la recherche, pour savoir s'il utilise les mêmes termes pour annoter et pour rechercher sur le web ? Les tests sont effectués

dans trois différents systèmes du Tagging : Delicious (page web), Flickr (images) et Last.fm (musique).

Dans **[Xu, 08]**, les auteurs mettent en œuvre le modèle vectoriel de la recherche d'information classique. En plus de la comparaison habituellement faite entre le vecteur requête et le vecteur contenu page web (Index), ils exploitent les annotations sociales (tags) d'un utilisateur comme étant son vecteur d'intérêt, qui va être comparé au vecteur 'Topic' (thème, titre) d'une page web qui n'est rien d'autres que les annotations assignées à cette page. C'est une recherche personnalisée à base d'annotations sociales.

Le rang d'une page est donc donné non seulement par le matching entre la page et la requête mais également par similarité entre utilisateurs et le 'Topic' de la page web.

Deux processus sont lancés :

- *Term matching process :* qui calcule la similarité entre requête et chaque page web pour générer une liste de documents sans prendre en compte l'utilisateur ;
- *Topic matching process :* qui calcule la similarité entre l'utilisateur (son vecteur d'intérêt) et chaque page web pour générer une liste de document avec prise en compte de l'utilisateur.

Une agrégation des rangs issus de chaque processus est effectuée selon la formule suivante :

$$R(u,q,p)=y.r_{term}(q,p)+(1-y).r_{topic}(u,p)$$

Formule 01 : *Calcul du rang d'une page web, [Xu, 08]*

Où $r_{term}(q,p)$ est le rang de la page p généré par le premier processus et $r_{topic}(u,p)$ est le rang de la page p généré par le deuxième processus, $0 \leq y \leq 1$, u l'utilisateur, q la requête et p la page web.

Les auteurs de **[Bao, 07]** proposent d'intégrer les annotations sociales (les tags) pour le calcul de similarité (Similarity Ranking) entre requête et document. En effet, l'ensemble des annotations que peut avoir une page, renseigne sur la popularité de celle-ci.

Ils proposent le SSR (SocialSimRank) qui mesure la similarité entre pages web et annotations sociales. Et le SPR(SocialPageRank) qui capture la popularité d'une page web du point de vues utilisateurs annotateurs, alors qu'habituellement la popularité d'une page est calculée uniquement du point de vue du créateur de la page et de l'utilisateur final (utilisateur des moteurs de recherche).

II.8. Rapprocher les ontologies et les folksonomies

Pour pallier les problèmes d'ambiguïté et de variation d'écriture cités précédemment, des études sont menées pour utiliser les ontologies dans les systèmes du Tagging. Ce recours aux ontologies est souvent appelé rapprochement d'ontologies aux folksonomies. Plusieurs travaux ont été effectués, **[Limpens, 08]** les classe en deux types, les approches d'extraction de liens sémantiques entre tags et les approches basées sur les ontologies comme support de folksonomies. Nous résumons ces deux classes de travaux comme suit :

II.8.1. Les approches d'extraction de liens sémantiques entre tags

Nous citons trois types d'approches, la première basée sur l'analyse des réseaux sociaux, la seconde analyse la dynamique des folksonomies et la dernière le clustering.

II.8.1.1. Analyse des réseaux sociaux appliquée aux folksonomies

Son objectif est de construire des ontologies légères (lightweight ontologies) à partir de l'analyse des folksonomies. **[Mika, 05]** propose un modèle 'tripartite' (ressource web, associées par un utilisateur (acteur) à une liste de concepts (tags)). Des méthodes d'analyse des réseaux sociaux (ARS) sont utilisées pour tisser des liens entre concepts et en déduire des catégories de termes ou des relations de subsumption (is a) :

1- Soit en rapprochant les tags ayant le plus de ressources en commun ; Ce cas sera un graphe concept-instance (ressource) reflétant la cooccurrence des tags sur les mêmes ressources (exemple : dans un site d'annotation de films, les tags *famille* et *enfant* ayant un certain nombre de ressources en commun, sont déclarés sémantiquement proche),

2- Soit les tags ayant plus d'acteurs en commun. Ce deuxième cas sera un graphe concept-auteur reflétant le regroupement par communauté d'intérêts (exemple : les utilisateurs ayant à la fois utilisé les tags sémantiquement proches *famille* et *enfant* forment une même communauté).

Le rapprochement entre tags ne peut se faire que si ces tags ont en commun plusieurs utilisateurs.

II.8.1.2. Analyse de la dynamique des folksonomies

La question posée est celle de la stabilité de distribution des tags pour une ressource donnée au fil du temps. **[Halpin, 07]** suppose que les tags les plus utilisés pour annoter une ressource demeurent les mêmes, et que la distribution de leur fréquence d'apparition suit une loi

de puissance. L'analyse montre que la distribution des tags pour les sites les plus populaires devient stable à partir d'un temps quasiment constant. Les auteurs se sont intéressés à la recherche de liens sémantiques entre les termes à l'aide de graphes de corrélations inter-tags.

II.8.1.3. Clustering

[Specia, 07] effectue un traitement statistique des annotations en regroupant les tags fortement liés entre eux en clusters (en utilisant une matrice de cooccurrence des tags sur les mêmes ressources), puis détailler les liens pouvant exister entre les tags du même cluster en cherchant dans des ontologies du web sémantique.

II.8.2. Les approches basées sur les ontologies

Les approches basées sur les ontologies comme support de folksonomies, soit elles utilisent l'ontologie pour guider le Tagging soit elles construisent des ontologies de folksonomies.

II.8.2.1. Guider le Tagging à l'aide d'ontologies

[Passant, 07] considère un tag comme propriété (HasTag) d'un concept d'une ontologie contrôlée. L'ontologie subordonne l'annotation en aiguillant le choix du tag. L'utilisateur est mené à réutiliser un tag existant, ou proposer un nouveau tag pour un concept existant ou proposer un concept et son tag s'il n'existe pas. Passant fait de chaque tag une propriété rattachée à des concepts dans l'ontologie d'une entreprise, cette approche permet, au cours de l'utilisation, de lever l'ambiguïté des tags, mais également d'enrichir l'ontologie lorsqu'un nouveau tag ne correspond à aucun concept existant.

II.8.2.2. Construire une ontologie de folksonomies

[**Gruber, 05]** suppose qu'il n'y a pas d'opposition entre les ontologies et les folksonomies et propose de construire une "ontologie de folksonomie".
La "TagOntology" en est un exemple, c'est un projet de construction d'ontologie pour la formalisation et la conceptualisation du Tagging. Elle met en œuvre quatre concepts : l'objet tagué (la ressource), le tag, l'utilisateur taguant et le domaine dans lequel le Tagging s'inscrit.

II.8.3. Exemples d'ontologies informatiques pour le Tagging

Plusieurs ontologies légères ont été conçues pour guider le processus du Tagging collaboratif, citons par exemple SIOC, SCOT, Common Tag, MOAT et NiceTag.

La figure Fig 06, schématise le modèle structurel de Common Tag, un tag donné pointe vers une autre ressource qui identifie le concept décrit par ce tag. Date et texte sont aussi des informations qui peuvent exister. Common Tag n'est pas encore une recommandation du W3C.

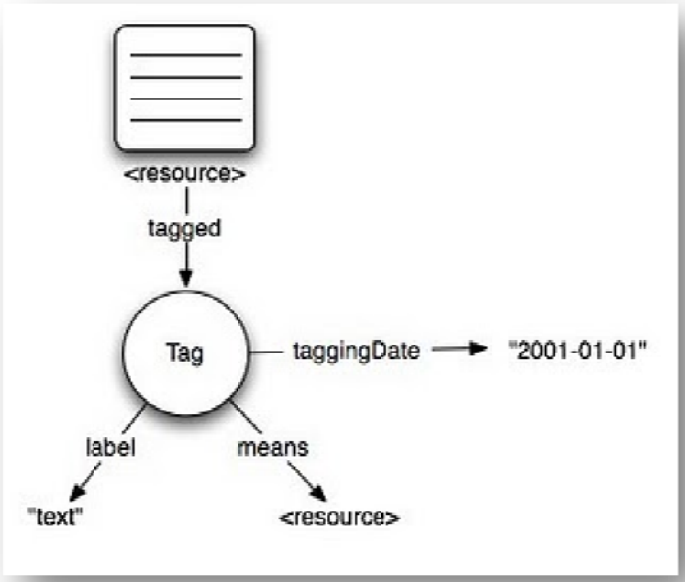

Fig 06 : *Schéma du modèle structurel de Common Tag*

[Passant, 08] dans son article sur les ontologies du web 2.0, parle de Moat (Meaning of a tag), et montre comment se déroule le processus du Tagging basé sur cette ontologie. Nous résumons ce processus dans la figure ci-dessous.

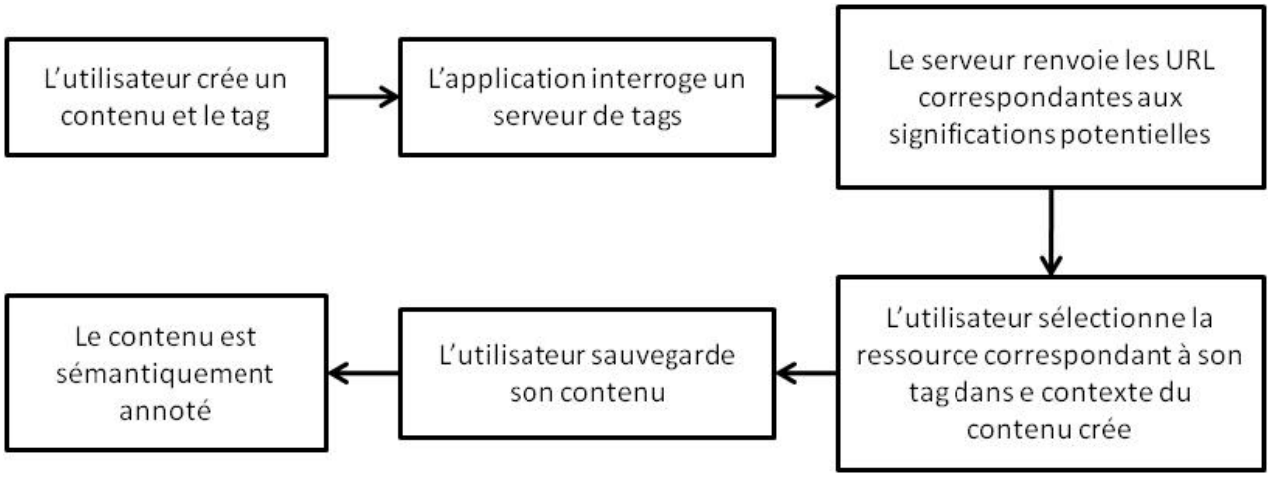

Fig 07 : *Processus de communication entre un client et un serveur MOAT*

La réalisation de l'action du Tagging à base d'ontologies (appelé Tagging sémantique), nécessite généralement la coopération de plusieurs ontologies, explique **[Passant, 08],** Une fois l'étiquetage effectué, une représentation RDF du contenu annoté et de l'action de *Tagging* (i.e.

la relation entre le contenu, l'utilisateur, le tag et sa signification) est produite. Cette modélisation se base sur différentes ontologies, à savoir SIOC pour la ressource elle-même, Tag Ontology pour la relation du *Tagging*, MOAT pour associer à cette relation la signification choisie, et à nouveau SIOC pour un lien direct entre ressource et signification. (Voir Fig 08).

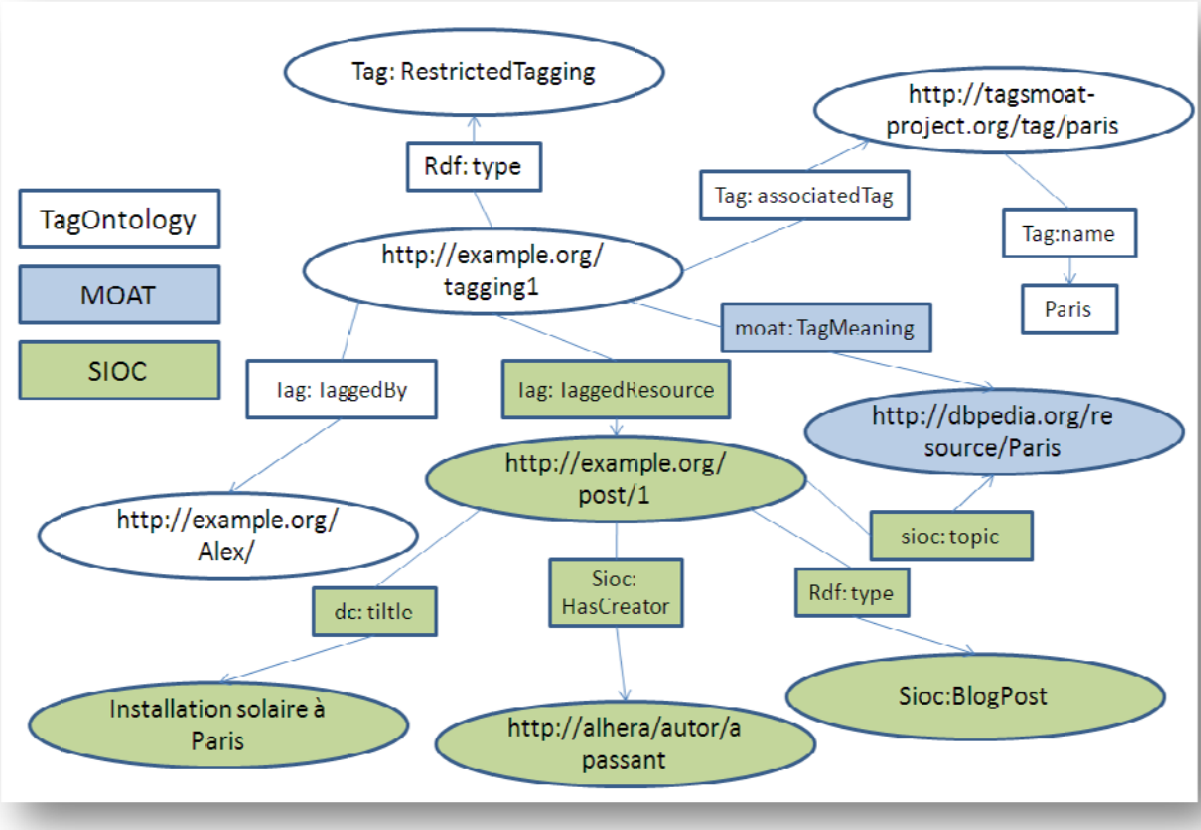

Fig 08 : *Action du Tagging pour un contenu donné. En blanc l'action elle-même (entre un utilisateur, une ressource et un tag), en vert le contenu et ses métadonnées, en bleu, l'URI de la signification du tag dans ce contexte,* ***[Passant, 08]****.*

II.9. Les limites des systèmes du Tagging

Nous mettons en évidence quelques remarques que nous avons recensées lors de l'utilisation de quelques systèmes du Tagging collaboratif :

Dans Delicious, l'un des systèmes les plus connus dans le Tagging, La valeur de la popularité d'un tag est le nombre de fois qu'il est cité pour tagguer l'ensemble des ressources. Le problème de variation d'écriture n'est pas résolu, exemple : les tags search_engine et searchengine sont utilisés pour tagguer la page www.google.com et sont considérés par le système comme deux

tags différents. Ces deux remarques sont valables pour pratiquement tous les autres systèmes. Un besoin en information de ressources tagguées avec *search_engine* n'inclura pas les ressources tagguées avec *searchengine,* (Silence).

Dans tous les systèmes de Tagging, aucun critère pour différencier les utilisateurs du point de vue de leurs maitrise et expertise dans le domaine de la ressource n'est pris en considération. Par conséquent, deux tags associés à une ressource respectivement par un utilisateur expert et un novice auront le même poids.

II.10. Conclusion

Entre avantages et inconvénients, le Tagging collaboratif semble susciter beaucoup d'intérêt, son application a atteint beaucoup de domaines tels que le e-learning et le développement du logiciel (soft developpement team) où sont exploités ses points forts à savoir la liberté des choix des termes et la collaboration. Nous avons, dans ce chapitre, donné la définition du Tagging et ses modèles. Nous avons parlé de folksonomie, ses avantages et inconvénients. Un petit comparatif entre la notion du Tagging et celle d'annotation a été établi. Nous avons cité une liste de travaux du domaine où nous avons mis l'accent sur ceux liés aux folksonomies et à la recherche d'information. Et nous avons terminé par citer un nombre de limites des systèmes du Tagging Collaboratif.

Etant un acteur important de ces systèmes, l'utilisateur qui crée les tags doit être étudié soigneusement et son rôle ainsi que sa représentation par le système mise en exergue. C'est ce qui va faire l'objet du chapitre suivant.

Chapitre III : Le Profil Utilisateur

Notre crime est d'être homme et de vouloir connaître.

Alphonse de Lamartine (Premières méditations poétiques)

III.1. Introduction

Le profil utilisateur est exploité pour différents buts. Dans le domaine de la recherche d'informations, il est utilisé pour raffiner les résultats des recherches et personnaliser celles-ci pour l'utilisateur. On parle donc de recherche personnalisée. Aussi il est exploité dans les systèmes de filtrages qui recommandent des informations qu'ils jugent intéressantes pour l'utilisateur.

Dans les systèmes du Tagging Collaboratif, lors de son inscription, l'utilisateur crée son profil. Sa constitution initiale diffère d'un système à un autre. La pluparts du temps, il est composé du nom, prénom, adresse électronique, mot de passe, et parfois une description personnelle.

Les informations initiales fournies par les utilisateurs sont statiques, seules, elles ne peuvent être très utiles pour la personnalisation de la recherche. L'aspect dynamique doit être pris en charge (par exemple changement de centres d'intérêts et évolution du profil). Pour cela, il existe des approches d'acquisition et d'enrichissement du profil, se basant sur les interactions de l'utilisateur avec le système telles que ses navigations, ses clicks…etc.

Le web 2.0 donne une nouvelle dimension pour l'acquisition du profil, c'est celle des tags. Ceux-ci peuvent être utiles pour l'extension et l'enrichissement du profil.

Dans ce chapitre, nous donnons les définitions du profil utilisateur, ses principales composantes (dimensions) et ses exploitations. Nous citons les approches existantes pour sa modélisation et son acquisition. Nous mettons l'accent sur l'apport des tags pour la définition du profil.

III.2. Définition

Selon **[Tamine, 06],** le profil de l'utilisateur est une structure d'informations hétérogènes qui couvre des aspects larges tels que l'environnement cognitif, social et

professionnel de l'utilisateur, ces informations sont généralement exploitées dans le but de préciser ses intentions au cours d'une session de recherche.

La notion de profil est apparue aux environs des années 80 avec les assistants et les agents d'interfaces, dû principalement au besoin de créer des applications personnalisées, capable de s'adapter à l'utilisateur **[Bouzeghoub, 05].**

Le profil utilisateur est caractérisé par un ensemble de traits concernant l'utilisateur : ses acquis (background), ses connaissances, ses objectifs (goals), ses préférences et son expérience dans un domaine donné, ses centres d'intérêts (interests) et ses traits individuels[1] **[Brusilovsky, 01]**.

III.3. Modélisation du Profil

Un des problèmes auxquels est confrontée l'utilisation du profil est sa modélisation. Comment mettre en évidence la diversité des informations nécessaires au filtrage d'informations, à la recherche d'information personnalisée ou à l'IHM adaptée. La modélisation du profil consiste à désigner une structure pour stocker toutes les informations qui caractérisent l'utilisateur et qui décrivent principalement ses centres d'intérêts en plus d'autres informations relatives à ses préférences, le contexte dans lequel il travaille, le but de ses recherches…etc **[Boulkrinet, 07].**

[Lechani, 05] fait une synthèse sur les modèles existants. Elle met en évidence l'approche canonique qui intègre des modèles utilisateurs typiques lors de la conception du système, cette approche présente l'inconvénient d'une description incertaine des situations. L'approche explicite caractérisée par le maintien d'une partie flexible contrôlée par l'utilisateur, elle présente l'inconvénient d'une surcharge cognitive pour l'utilisateur. Enfin l'approche automatique qui vient remédier aux inconvénients des deux premières approches. Dans cette approche on infère le modèle utilisateur de manière implicite à partir des données collectées durant ses utilisations du système. L'approche automatique est la plus répandue, sa mise en œuvre est faite par deux principales classes de techniques : collaboratives et statistiques.

[1] Les traits individuels de l'utilisateur sont ses caractéristiques personnelles telles que : introverti, extraverti et les facteurs cognitifs qui sont généralement tirés par test psychologiques, [Boulkrinet, 07].

III.4. Représentation du Profil

La représentation de l'utilisateur à travers le profil permet de mieux comprendre ses mécanismes cognitifs. **[Lechani, 05]** cite cinq types de représentation ou modélisation du profil utilisateur :

III.4.1. Représentation ensembliste ou vectorielle

Inspirée du modèle classique de Salton, le profil est alors représenté par un ou plusieurs vecteurs définis dans un espace de termes et dont les coordonnées correspondent à leurs poids respectifs. L'utilisation de plusieurs vecteurs permet la prise en compte de la diversité des centres d'intérêts. Ce type de représentation offre l'avantage de la simplicité de sa mise en œuvre.

III.4.2. Représentation sémantique

Cette représentation met en évidence les relations sémantiques entre les informations que contient le profil, elle est basée sur l'utilisation d'ontologie.

III.4.3. Représentation connexionniste

C'est un type de représentation basé sur l'interconnexion de nœuds représentant les termes qui composent le profil.

III.4.4. Représentation multidimensionnelle

Le profil est structuré selon un ensemble de dimensions et celles-ci en sous dimensions.

III.4.5. Représentation hiérarchique

Dans ce cas le profil utilisateur est représenté à travers la construction d'une hiérarchie de concepts au lieu d'un ensemble de domaines indépendants. Chaque catégorie de la hiérarchie représente la connaissance d'un domaine d'intérêts de l'utilisateur, la relation généralisation/spécification entre les éléments de la hiérarchie traduit d'une manière plus réaliste les centres d'intérêts de l'utilisateur qui ne sont pas toujours indépendants les uns des autres.

III.5. Dimensions d'un Profil utilisateur

Le contenu d'un profil utilisateur diffère d'un système à un autre et d'un domaine à un autre.

Dans le domaine des IHMs, le profil contient des informations permettant au système d'adapter l'affichage des résultats selon les préférences de l'utilisateur. En plus des informations sur l'identité (nom, prénom…), il contient d'autres informations telles que ces centres d'intérêts, le contenu de certaines catégories d'un centre d'intérêt peut être déduit à partir des données personnelles (exemple : son horoscope à partir de sa date de naissance). Dans le domaine de la RI, le profil décrit le plus souvent les centres d'intérêts de l'utilisateur, et représenté sous forme de vecteur de mots clés pondérés.

En se basant sur le modèle générique du profil qu'a défini **[Amato, 99]**, et repris dans **[Bouzghoub, 05]** nous citons les huit principales catégories :

- **Les données personnelles :** cette catégorie comprend les données d'identification de l'utilisateur, à savoir le nom, prénom, date et lieu de naissance, sexe, login et mot de passe. Peut aussi contenir ses coordonnées personnelles et professionnelles. Cette catégorie représente la partie statique du profil.
- **Les centres d'intérêt :** Le centre d'intérêt exprime le domaine d'expertise de l'utilisateur ou son périmètre d'exploration. Il peut être défini par un ensemble de mots clés ou un ensemble d'expressions logiques (requêtes). Toute requête émise par l'utilisateur sera enrichie avec les mots clés ou les prédicats des requêtes définissant le centre d'intérêt.
- **L'ontologie du domaine :** elle permet de définir la sémantique de certains termes ou opérateurs employés par l'utilisateur notamment dans sa description de ses centres d'intérêts.
- **La qualité attendue :** cette catégorie concerne les préférences de l'utilisateur quant aux documents qu'il cherche. Des préférences liées au contenu du document (exemple thème et langue du document). D'autres liées aux structures de documents comme le format (texte, image…), le type (article, proceeding, page web…). Enfin des préférences liées aux sources des documents (par exemple : des URLs, auteurs et éditeurs spécifiques).
- **La Customisation :** concerne ce qui est lié aux modalités de présentation des résultats en fonction de la plateforme, de la nature et du volume des informations. Elle comprend aussi les préférences esthétiques ou visuelles de l'utilisateur.

- **Les actions de l'utilisateur (Feedback) :** c'est un ensemble d'informations explicites, fournies par l'utilisateur lui-même (exemple : jugements sur la pertinence des résultats), ou implicite, déduites par le système à partir de certaines actions de l'utilisateur (exemple : page web visitées, documents lus, sauvegardés…). Ces informations servent à enrichir et étendre le profil.
- **La sécurité et la confidentialité :** elle permet à l'utilisateur d'exprimer les traitements ou actions qu'il souhaite être privées, ou à accès restreint. Ces conditions d'accès exprimées par l'utilisateur peuvent concerner toutes les catégories précédentes.
- **Informations diverses :** varient selon les applications, cette catégorie comprend les informations qui ne peuvent être incluses dans aucune des catégories précédentes.

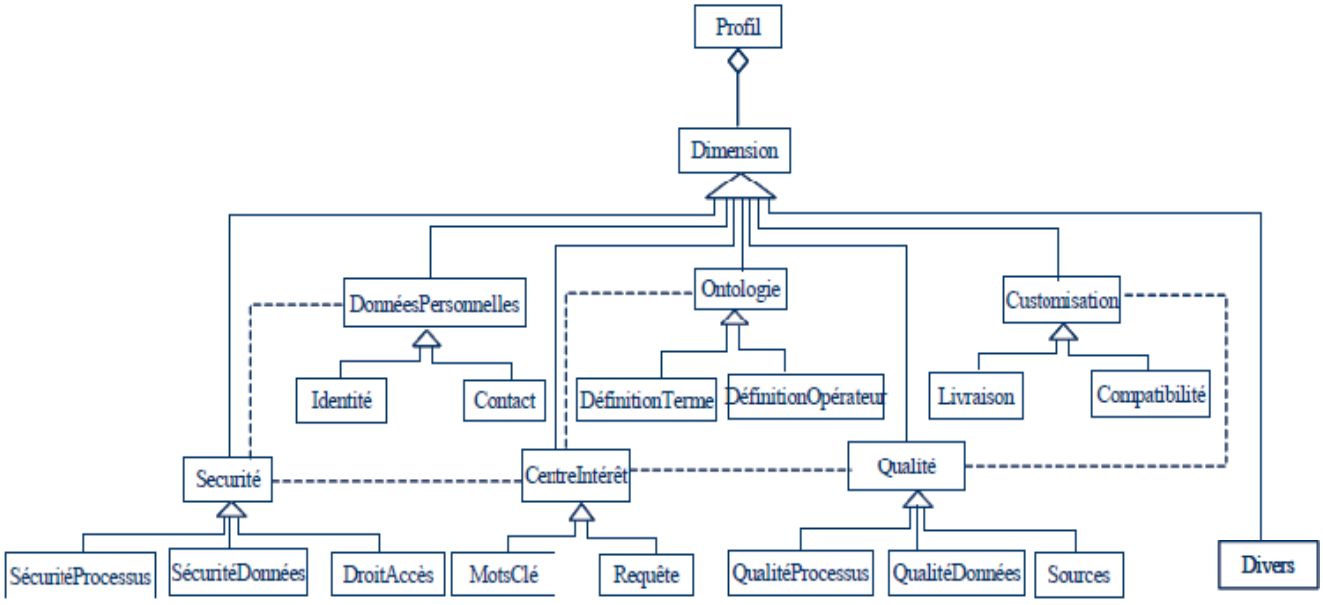

Fig 09 : *Dimensions et sous-dimensions d'un modèle du profil,* ***[Bouzghoub, 05]***

Les dimensions et sous-dimensions définissant un profil ne sont pas indépendantes les unes des autres ; elles peuvent être liées par des associations sémantiques qui caractérisent leurs dépendances ou leurs corrélations. **[Bouzghoub, 05].**

La définition d'un profil d'un utilisateur particulier pour une application donnée revient à sélectionner les dimensions jugées utiles de ce schéma (Fig 09), à les instancier et à en dériver un sous-schéma qui définira le profil de cet utilisateur.

III.6. Acquisition du Profil

La manière la plus simple d'acquérir un profil est celle qui se base sur les informations qu'introduit l'utilisateur lui-même. Or, ce n'est pas toujours évident d'avoir les informations

cruciales, notamment pour le filtrage. Plusieurs approches ont été proposées pour acquérir et construire un profil utilisateur. **[Boulkrinet, 07].**

III.6.1. Approche simpliste :

C'est une approche qui préconise l'introduction et la description du profil par l'utilisateur. Elle est simple certes, mais son principal inconvénient est l'effort demandé à l'utilisateur et qui peut par fois être énorme.

III.6.2. Approche dynamique :

Cette approche permet la construction dynamique du profil en se basant sur ses interactions avec le système, pour réajuster le profil initial (ensemble de mots clés fourni par l'utilisateur).

III.6.3. Approche par apprentissage :

L'idée est d'observer le comportement de l'utilisateur à travers ses interactions, ses clicks, ses navigations, lectures et consultations pour apprendre son profil et prévoir son comportement futur.

III.7. Modèle conceptuel du profil

Se basant sur les dimensions de **[Amato, 99]**, **[Bouzeghoub 05]** a proposé un modèle conceptuel, multidimensionnel, dans l'objectif de proposer un ensemble de dimensions ouvertes, capables d'accueillir la plupart des informations caractérisant un profil (Fig 10).

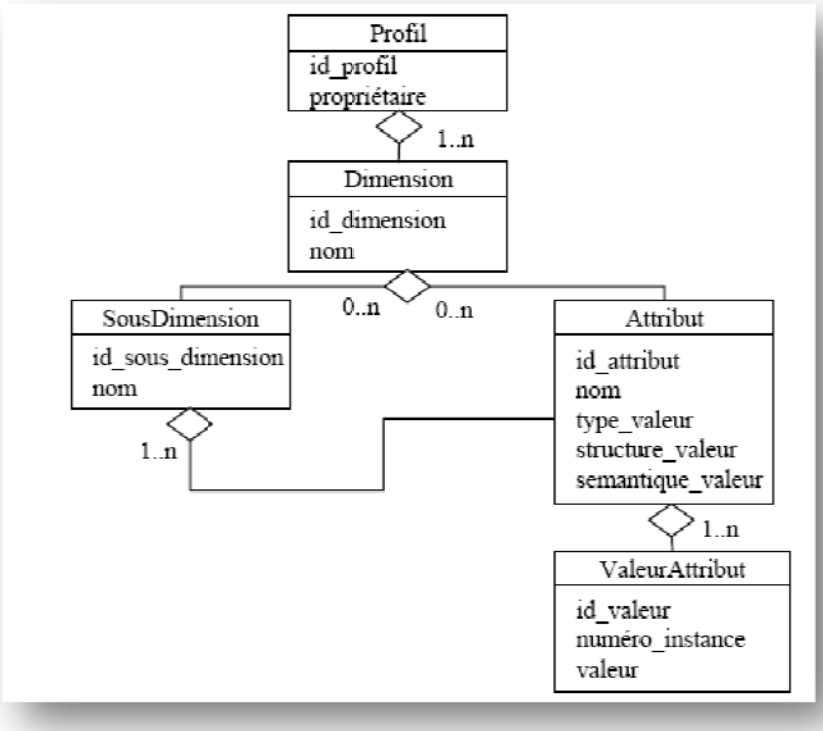

Fig 10 : *Modèle conceptuel du profil utilisateur,* ***[Bouzeghoub, 05]***

Chaque dimension est constituée d'un ensemble d'attributs dont les valeurs peuvent être simples (valeur numérique ou symbolique) ou complexes (expression logique, fonction d'utilité ou ordre de préférence par exemple). Certaines dimensions sont organisées en sous-dimensions. Un attribut du profil est défini par un identificateur, un nom, un type, une expression de préférence et une sémantique. Le type peut être entier, réel, chaine de caractères, ensemble…, structure_valeur est une expression de préférence qui peut être un vecteur de termes pondérés, une expression logique…, la sémantique d'un attribut peut être définie à l'aide d'une ontologie.

III.8. Profil utilisateur dans le Web 2.0

Une nouvelle information vient enrichir le profil utilisateur, c'est les annotations sociales (tags), à laquelle s'intéressent les chercheurs pour améliorer les systèmes de recommandation et filtrage d'informations d'une part et d'autre part exploiter le profil utilisateur pour améliorer la recherche à base de tags notamment dans les systèmes du Tagging collaboratif.

III.8.1. Recherche à base de tags et profil utilisateur

La principale action dans les systèmes du Tagging collaboratif est bien entendu l'association des tags aux ressources par les utilisateurs, cette action permet une recherche et une exploration de contenus à base de ces tags.

Plusieurs travaux sont menés pour améliorer la recherche à base de tags. Par exemple, en se basant sur le fait que l'exploration basée uniquement sur la popularité est limitée, **[Wang, 09]** suggère de proposer à l'utilisateur un nuage de tags spécifique et propose d'incorporer le profil utilisateur pour classer les ressources par leurs degrés de pertinence en se basant sur un modèle probabiliste dans le processus de recherche. D'autres travaux tentent d'exploiter ce profil utilisateur de manières différentes, tels que **[Xu, 08],** cité dans la section 7 du chapitre précédent, qui met en œuvre le modèle vectoriel de la recherche d'information classique, et introduit les tags d'un utilisateur comme étant son vecteur d'intérêt. **[Bao, 07]** propose d'intégrer les tags pour le calcul de similarité (Similarity Ranking) entre requête et page web.

III.8.2. Créer et enrichir le profil utilisateur en se basant sur les tags

Le profil utilisateur est crée au moment de l'inscription de l'utilisateur, il est généralement constitué d'un identifiant, mot de passe et quelques données personnelles (nom, prénom, sexe…). Les systèmes du Tagging collaboratif actuels exploitent le profil uniquement pour l'authentification et permettent ainsi aux utilisateurs de visualiser leurs propres ressources, bookmarks, tags et ceux des autres utilisateurs. Or ces systèmes peuvent tirer profit des différentes annotations accomplies par les utilisateurs pour enrichir leurs profils.

[Carmagnola, 08] considère que les tags sont un nouveau type de feedback (retour) de l'utilisateur, et peut être un indicateur très important sur les préférences de celui-ci. Les actions du Tagging fournissent ainsi des informations pouvant être utilisées pour améliorer les connaissances du système sur l'utilisateur.

Différentes approches de construction du profil basées sur le Tagging sont présentées dans **[Cayzer, 09].** Les deux approches les plus utilisées pour la construction des centres d'intérêts des utilisateurs à partir de leurs tags sont l'approche naïve et l'approche par cooccurrence :

- ***L'approche naïve :*** La méthode la plus simple de construire un profil, est de compter les occurrences des tags, le résultat est une liste de tags classés par ordre de popularité. Cette approche construit donc le profil avec le top des tags cités par l'utilisateur pour tagguer l'ensemble des ressources. Sa simplicité et sa rapidité de mise en œuvre font d'elle une approche très utilisée notamment sous la forme de

nuages de tags, cependant les tags résultants sont généralement des termes génériques et sont sélectionnés au détriment des termes spécifiques.

- ***L'approche par cooccurrence :*** Cette approche consiste en la création d'un graphe où les nœuds représentent les tags cités par l'utilisateur et les arcs les relations de cooccurrence entre ces tags. Les arcs sont pondérés par le nombre de cooccurrences. Le profil résultant est le top k des nœuds participants aux arcs ayant les plus grands poids. Cette approche est largement utilisée pour la détection des relations entre tags **[Cattuto, 07]**, l'extraction d'ontologies légères **[Mika, 05]**, et la recommandation de tags **[Xu, 06]**. Elle permet la construction d'un profil plus précis que celui obtenu avec l'approche naïve. Cependant, elle présente l'inconvénient de négliger les ressources à tag unique.

[Rupert, 10] regroupe les ressources en clusters et les utilisateurs en communautés en créant leurs profils. Puis associe ces ressources à une communauté d'utilisateurs en se basant sur les systèmes multi-agents. Son modèle est utilisé pour un système de recommandation de ressources.

[Firan, 07] a crée le profil utilisateur à base des tags utilisés pour une meilleure recommandation de musique sur Last.fm. Le profil est calculé comme suit :

$$Profil_{\_Tags(U)} = \{< TGi, Pi > | TGi = les\ tags\ de\ l'utilisateur\ U, Pi = p(TGi, U)\}$$

Formule 02 : *Formule de création du profil,* ***[Firan, 07]***

Avec $p(TGi, U)$ la préférence de l'utilisateur U pour le tag TGi. La notion de préférence est calculée par **[Firan, 07]** avec une fonction logarithme selon que l'utilisateur écoute la musique via le site last.fm ou via les morceaux existants sur son ordinateur.

[Liang, 09] donne une composition différente du profil utilisateur à base des tags, pour le même but que les deux auteurs précédents, améliorer un système de recommandation, il définit le profil de la manière suivante :

UFi = (Tu_i, Pu_i, TP_i)

avec :

- Tu_i={t_j|t_j∈T, ∃p_k∈P, E(u_i,t_j, p_k) =1} l'ensemble des tags de l'utilisateur u_i, où p_k est un ensemble de ressources, E(u_i,t_j, p_k) une fonction désignant que l'utilisateur u_i a associé t_j pour p_k .

- P_{ui}={pk|pk∈T, ∃tj∈P, E(ui,tj, pk) =1} l'ensemble des ressources que l'utilisateur ui a taggué.
- TPi={<tj, pk>| tj∈T, pk∈P, and E(ui,tj,pk)=1} la relation entre les tags de l'utilisateur et les ressources.

En vue de créer un profil utilisateur précis et dynamique à base des tags, **[Huang, 08]** dans son article 'You are what you tag', a introduit la notion de capacité du tag à représenter une ressource en se basant sur deux facteurs, l'ordre du Tagging et la popularité. Selon **[Golder, 05]** le premier tag cité par un utilisateur pour une ressource est plus représentatif que les suivants, le poids basé sur l'ordre est donné dans **[Huang, 08]** par la formule suivante :

$$W_{order}(t_i,p) = \frac{1}{|T(c,p)|} \times \begin{cases} exp^{-1/10} \; if \; i \leq 10 \\ exp^{-1} \; if \; i > 10 \end{cases}$$

Formule 03 : *Formule du calcul du poids d'un tag basé sur l'ordre,* ***[Huang, 08]***

Avec ti le tag d'ordre i, p l'utilisateur ayant associé T(c; p) = {t1; t2; ….. ; tn} pour le contenu c.

Une somme des poids w_{order} de tous les utilisateurs est calculée comme suit :

$$capacity(t,c) = \frac{1}{|P(c)|} \times \sum_{p \in P(c,t)} W_{order}(t,p)$$

Formule 04 : *Formule du calcul de la capacité d'un tag,* ***[Huang, 08]***

Avec p(c,t) est l'ensemble des utilisateurs ayant associé t pour c, et |P(c)| le nombre d'utilisateurs ayant taggué le contenu c.

III.9. Conclusion

Nous avons vu dans ce chapitre les différentes notions liées au profil utilisateur, ses modèles, ses différentes représentations, son acquisition et ses dimensions. Nous avons évoqué la nouvelle donnée sociale qui pourrait enrichir d'avantage le profil utilisateur et qui est tant exploitée dans le domaine de la recherche d'information personnalisée, et nous avons cité quelques travaux du domaine.

Dans le cadre de notre travail et dans la même idée de **[Huang, 08],** nous exploitons le profil utilisateur pour procéder à une pondération de tags, afin de les classer et filtrer les plus représentatifs. Pour cela nous adoptons un modèle utilisateur permettant de contenir ses informations cruciales et une méthode de construction de celui-ci.

Dans le chapitre suivant, nous présentons une nouvelle approche dont l'objectif est de construire un ensemble de tags décrivant les ressources de la manière la plus précise et la plus exacte possible. Ce descripteur sera utilisé pour des besoins de recherche ou de classification. L'idée est donc, d'exploiter le profil utilisateur comme critère de classement des tags en plus de la popularité.

Deuxième Partie
Une Approche de Filtrage de Tags à base du Profil utilisateur

Chapitre IV : Présentation de l'approche

La réflexion jointe à l'usage donne des idées nettes

(Jean-Jacques Rousseau Confess. V)

IV.1. Introduction

Les systèmes du Tagging Collaboratif actuels 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 du plus populaire au moins populaire.

Citons l'exemple de Delicious et Technorati, les tags les plus représentatifs d'une ressource sont les plus populaires (les plus cités par les utilisateurs). La popularité est calculée en considérant un même poids pour tous les utilisateurs. Un tag T1, associé à une ressource R1 et cité par les utilisateurs U1, U3, U4, U10, a un score de popularité de quatre (04). Aucune différenciation n'est faite entre les utilisateurs, et un classement décroissant de popularité est établi par le système, pour permettre un accès direct à la ressource via ces tags populaires. Cependant, d'autres paramètres peuvent contribuer à un meilleur classement. En effet, ces tags populaires sont ils objectifs ? Représentent-ils au mieux le contenu qu'ils annotent ? Donnons l'exemple d'un utilisateur qui attribue le tag 'Beau-paysage' pour une image d'un paysage et un autre qui attribue le tag 'Haïti'. Ces deux tags n'ont pas la même efficacité (ou capacité) à représenter l'image, le premier est très générique, le deuxième et précis. Un troisième 'Cap-Haïtien' aurait pu être plus précis.

L'idée de notre travail est de trouver un autre critère de pondération des tags que celui de la popularité. Notre approche consiste à intégrer l'utilisateur dans le calcul du poids du tag associé à une ressource donnée. Pour cela, nous présentons en premier lieu, nos motivations, le principe général de l'approche. Dans un second lieu, nous présentons un modèle du profil utilisateur permettant de contenir ses informations personnelles, son activité et son expertise, ainsi qu'une méthode de construction de ce profil à base du Tagging. Ensuite, nous proposons une formule de pondération des tags intégrant le profil utilisateur. Nous terminons par une conclusion et des perspectives.

IV.2. Motivations

Dans un système de Tagging collaboratif, et avec l'ampleur que prend ce type de système, des centaines de nouveaux utilisateurs s'inscrivent chaque jour. Une ressource donnée peut être tagguée par cette multitude d'utilisateurs, de plus la notion d'adéquation de tag est perçue dans les systèmes actuels du point de vue popularité uniquement. L'expertise de l'utilisateur n'étant pas prise en compte, les tags de ce dernier ont toujours le même poids qu'il soit novice ou expert. D'un autre côté, l'utilisateur se perd souvent dans une multitude de tags décrivant la ressource. Vu que ces derniers sont, le plus souvent, présentés en globalité sans filtrage ni limitation du nombre.

Tout cela peut être source de certains problèmes que nous résumons dans le tableau ci-dessous.

Problème	Conséquence
Nombre énorme de tags pour une même ressource	Risque de présence de tags inadéquats et donc non satisfaction d'un besoin utilisateur (bruit).
Utilisateur novice	Tags non appropriés au contenu de l'objet. Problème du bruit.
Tags non populaire mais issus d'un utilisateur connaisseur	Problème du silence.

Tableau 04 : *Problèmes dus à l'utilisation unique de la popularité*

IV.3 Principe général

L'objectif de notre approche est de construire un ensemble de tags décrivant les ressources de la manière la plus précise et la plus exacte possible. L'idée est de trouver un autre critère de classer les tags, ceux-ci classés actuellement uniquement par leur popularité, sans aucune différenciation entre les utilisateurs.

L'approche consiste à intégrer l'utilisateur dans le calcul du poids du tag associé à une ressource donnée. Pour cela, nous présentons en premier lieu, un modèle du profil utilisateur permettant de contenir son activité et son expertise, ainsi qu'une méthode de construction de ce profil à base du Tagging. Ensuite, nous proposons une formule de pondération des tags intégrant le profil utilisateur. Les tags pondérés ainsi obtenus sont classés par ordre décroissant de leur poids et les n premiers forment le descripteur de la ressource (Fig 11).

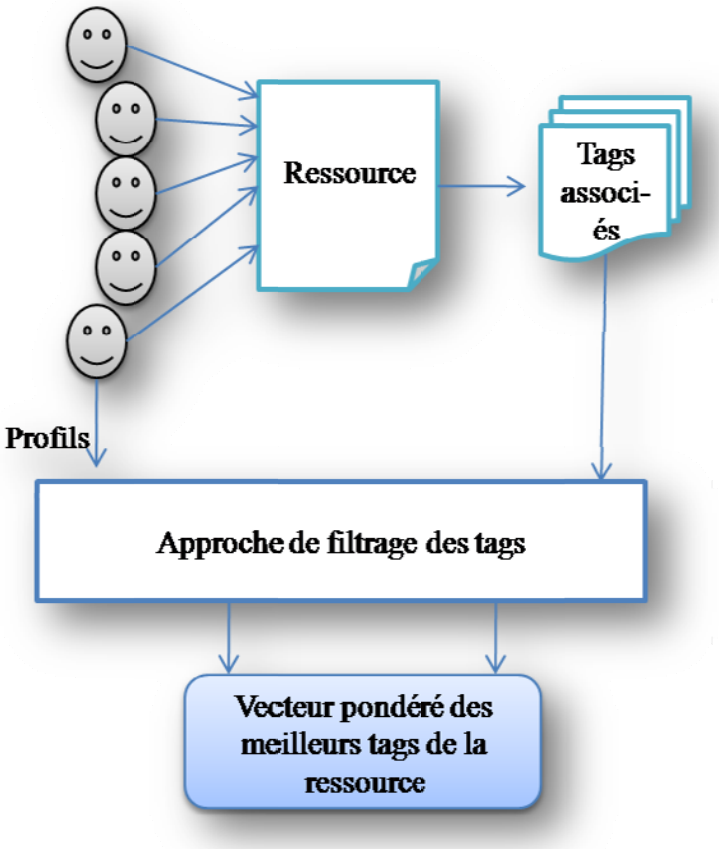

Fig11 : *Principe général de l'approche*

IV.4 Présentation de l'approche

L'enchainement global de notre approche consiste en trois étapes illustrées dans la figure ci-dessous (Fig12):

- Modélisation et construction du profil utilisateur ;
- Pondération et classement des tags ;
- Création des descripteurs de ressources.

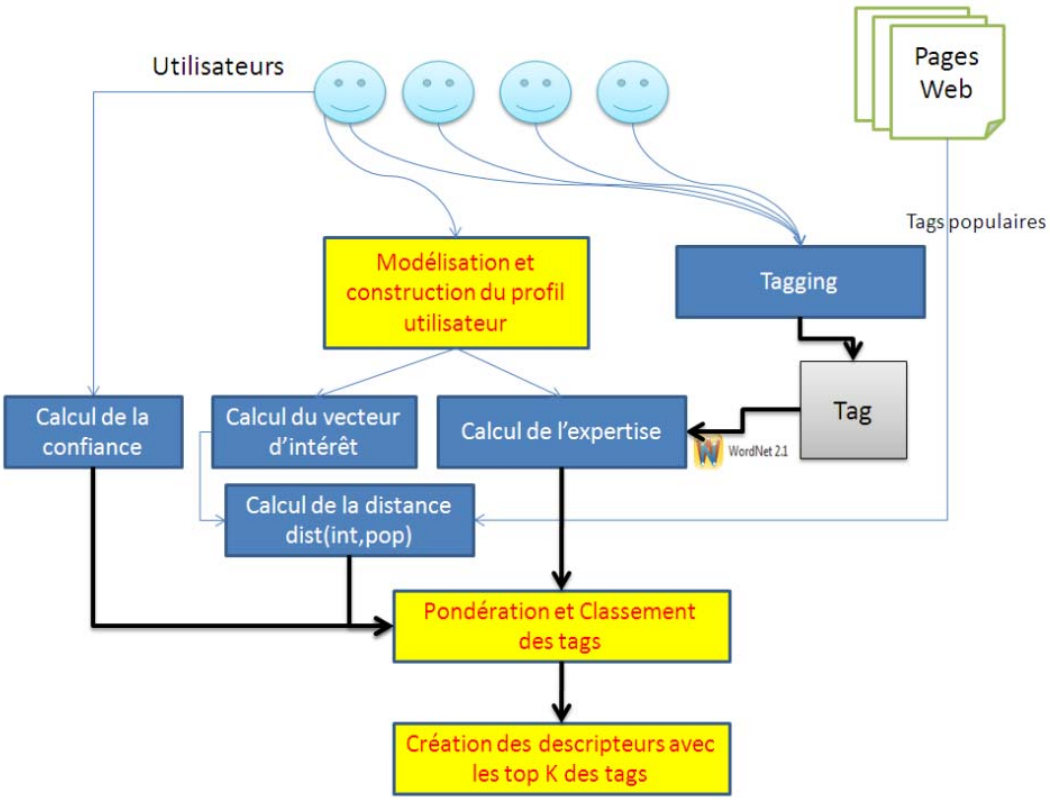

Fig12 : *Schéma global de l'approche*

La première étape consiste à proposer une modélisation du profil et construire ses dimensions. La seconde étape concerne la pondération des tags en prenant en compte les utilisateurs et le classement de ceux-ci. Enfin, la troisième étape consiste à construire un descripteur de ressource composé des meilleurs tags (filtrage). **[Kichou, 11]**.

La figure suivante (Fig13) présente le processus détaillé de l'approche dont les différentes actions sont détaillées tout au long de ce chapitre.

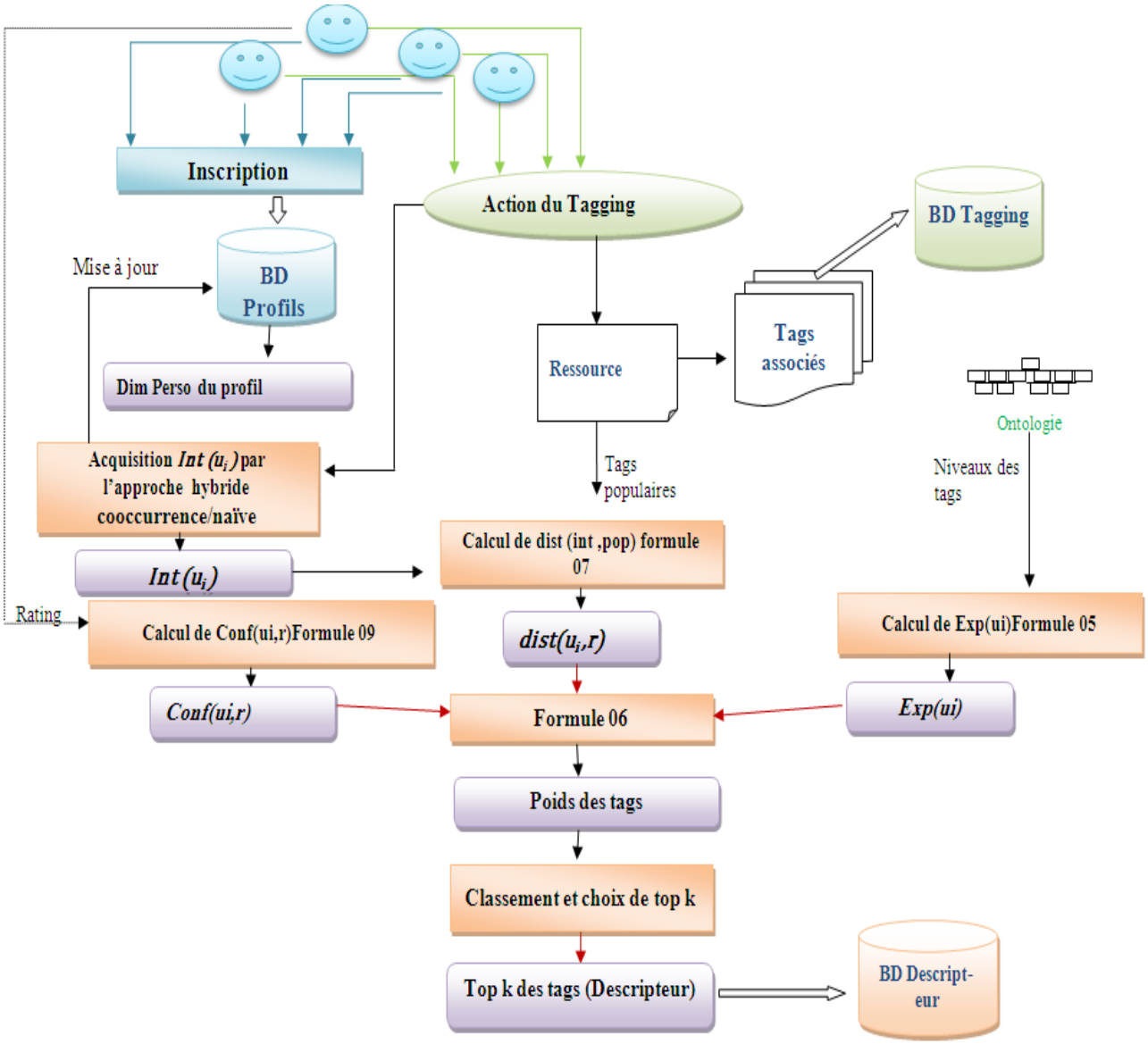

Fig 13 : *Processus détaillé de l'approche*

IV.4.1. Le modèle du profil utilisateur

Pour intégrer le profil utilisateur dans le calcul des poids des tags, nous définissons un modèle utilisateur représentant des informations reflétant son activité dans le système. Nous décrivons le modèle de l'utilisateur adopté puis nous expliquons la démarche de construction de celui-ci.

IV.4.1.1. Représentation du profil

La définition du profil utilisateur pour une application donnée revient à sélectionner les dimensions jugées utiles **[Bouzghoub, 05]** (voir section 5 du chapitre III). Dans notre cas, un utilisateur est défini par trois dimensions. La première contenant ses informations

personnelles, la seconde représente ses centres d'intérêts et la dernière nous renseigne sur son degré d'expertise dans le domaine (Fig14).

IV.4.1.1.1. La dimension personnelle

Sert à identifier l'utilisateur, (identifiant, nom, prénom, login, mot de passe…). Ces informations sont introduites par l'utilisateur lors de son inscription dans le système.

IV.4.1.1.2. La dimension centres d'Intérêts

Une ressource dont le contexte est proche du domaine d'intérêt de l'utilisateur a beaucoup de chance d'être tagguée de manière plus efficace. La dimension centre d'intérêts Int(ui) nous renseigne sur les intérêts et préférences de l'utilisateur. Elle est représentée sous forme de vecteur de tags pondérés, construit en utilisant la combinaison de deux approches de construction de profil, l'approche naïve **[Cayzer, 09]**, et l'approche par cooccurrence issue des techniques de l'analyse des réseaux sociaux **[Wasserman , 94]**. Int(ui) = {(t1,w1), (t2,w2)……(tj,wj)}avec t_i le tag d'indice i et w_i le poids de ce tag.

Exemple : Int(ui) = {(Web,5), (Delphi, 3), (Java, 2), (Programming, 2)}.

IV.4.1.1.3.La dimension expertise

Les utilisateurs experts dans un domaine, ont tendance à utiliser des termes spécifiques pour tagguer vu qu'ils ont une parfaite maîtrise des concepts de ce domaine. Cette dimension est le degré de maitrise de l'utilisateur du domaine des ressources tagguées. Elle est fonction du niveau des tags cités par l'utilisateur dans l'ontologie du domaine utilisée à cet effet. Plus l'expertise est importante, plus l'utilisateur est proche du contexte de la ressource.

Exemple : Un utilisateur ayant cité les tags 'eclipse', 'javascript', 'html' est considéré plus expert qu'un autre utilisateur ayant cité les tags 'language', 'programming' dans le domaine de l'informatique.

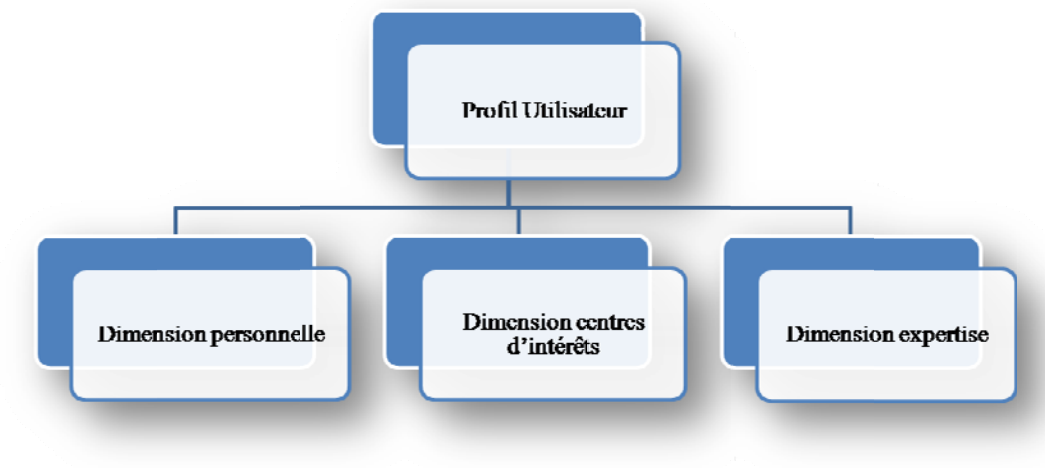

Fig14 : *Dimensions du Profil utilisateur défini dans l'approche*

IV.4.1.2. Construction du profil

La construction du profil utilisateur revient à construire les dimensions Int(ui) et Exp(ui) en se basant sur les opérations du Tagging qu'il effectue.

IV.4.1.2.1. Construction de la dimension centres d'intérêts

Il existe plusieurs approches de construction d'intérêts des utilisateurs (section 8.2 du chapitre III), nous présentons ci-dessous, une proposition de combiner les deux approches les plus utilisées, en l'occurrence l'approche naïve et par cooccurrence.

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

Chacune des approches naïve et par cooccurrence présente des inconvénients, nous proposons de combiner ces deux approches. Nous estimons que la combinaison de ces dernières, non seulement 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 de l'approche naïve et l'approche par cooccurrence est un graphe de nœuds et d'arcs pondérés (Fig15). Les nœuds (tags) pondérés appartenant aux arcs ayant les plus grands poids composent notre vecteur Intérêts.

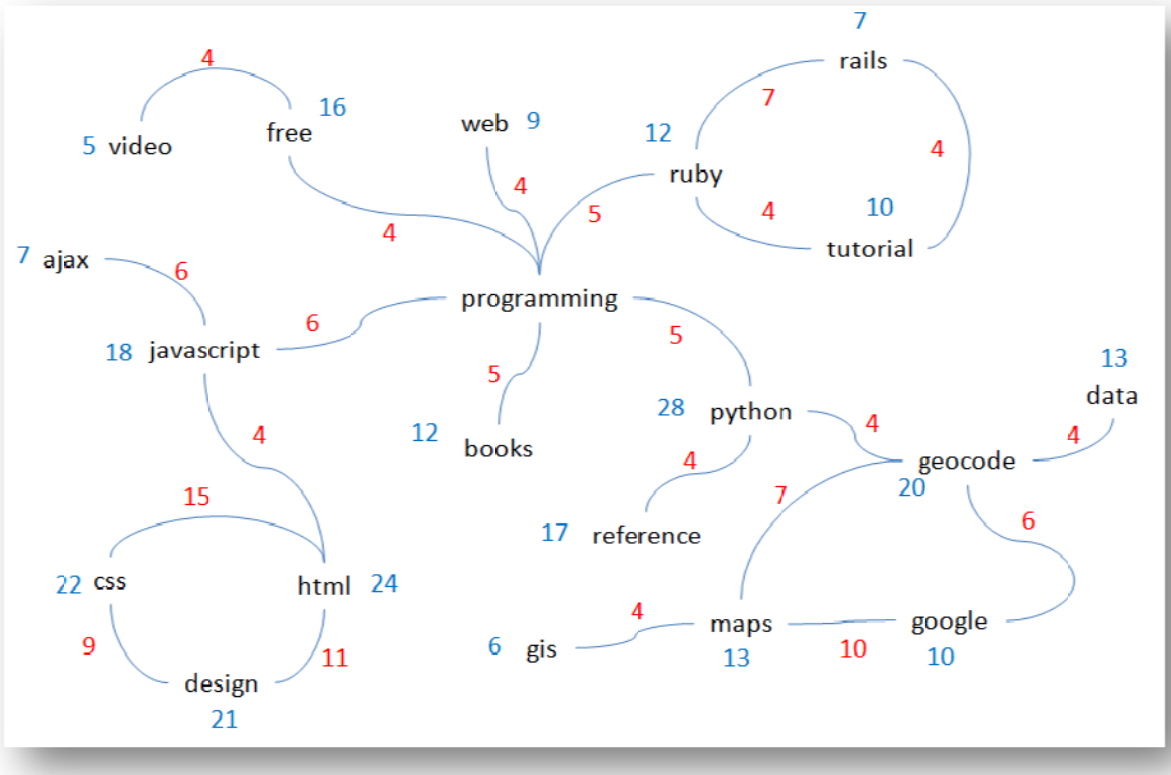

Fig15: *Exemple de graphe construit avec la combinaison de l'approche naïve et l'approche par cooccurrence*

La figure Fig15 illustre le graphe du profil d'un utilisateur avec les poids de cooccurrences des tags ainsi que la popularité de chacun d'eux. La figure ci-dessous présente le vecteur d'intérêt de cet utilisateur selon l'approche naïve, par cooccurrence et l'approche hybride avec k=5.

Naive	37 programming	28 python	24 html	22 css	21 design
Cooccurrence	css	html	design	google	maps
Hybride	22 css	24 html	21 design	10 google	13 maps

Fig16 : *Comparaison des vecteurs intérêts construits avec les différentes approches*

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

L'algorithme Add-A-Tag a été proposé dans **[cayzer, 09]** pour mettre en œuvre une nouvelle approche de construction de profil, 'Adaptive approach'[2] . nous adaptons cet algorithme de manière à réaliser l'approche hybride.

Soit u un utilisateur taggant un nombre de ressources avec l'ensemble de tags : $T=\{t_1,t_2,\ldots t_n\}$. Le graphe du profil de l'utilisateur u G_u (V, E) où $V=\{v_1,v_2,\ldots..v_n\}$ est l'ensemble des nœuds (Vertices), et $E=\{e_1,e_2,\ldots e_n\}$ est l'ensemble des arcs (Edges).

Etape 1 : Mise à jour du graphe

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

1. Pour chaque tag t_x avec x ϵ 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 nœud du graphe son poids (sa popularité).

Etape 2 : Extraction du profil

1. Créer un sous ensemble E_s de E, ordonné avec un ordre décroissant des poids des arcs ;
2. Choisir le top k des éléments de E_s avec k un entier non nul (k>0) ;
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.

IV.4.1.2.2. Construction de la dimension expertise

Un utilisateur expert dans un domaine, a une parfaite maitrise des termes spécifiques de ce domaine. Il a donc tendance à associer ces termes spécifiques aux ressources qu'il taggue, en pharmacie par exemple, un expert associe le nom de la molécule d'un médicament, alors qu'un novice se contente d'associer le terme 'médicament'. Dans notre approche, nous

[2]L'approche adaptative est une extension de l'approche par cooccurrence, dans laquelle **[cayzer, 09]** introduit la notion d'âge du tag pour favoriser les nouveaux tags associés par l'utilisateur.

utilisons une ontologie et nous observons les tags associés par l'utilisateur pour l'ensemble des ressources, et situer leurs niveaux (profondeurs) dans la hiérarchie de l'ontologie. Plus le tag utilisé est profond (vers les feuilles) plus l'utilisateur est expert. L'expertise est la moyenne des profondeurs des tags de l'utilisateur, calculée comme suit :

$$Exp(ui) = \frac{\sum prof(tj)}{|Tu|}$$

Formule 05 : *Formule du calcul de l'Expertise*

Où Prof(tj), profondeur du tag tj, est le nombre de nœuds le séparant de la racine ; Tu est un sous-ensemble de la personomie[3] de l'utilisateur ui contenant les tags que celui-ci a associé aux ressources, définie comme suit :

Tu= {tj| (ui, tj, r) ∈ Y} avec Y l'ensemble des annotations (actions du Tagging).

- **Choix de l'ontologie**

Afin de localiser les niveaux des tags utilisés pour le calcul de l'expertise, nous avons choisi d'utiliser WordNet. C'est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton[4]. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. L'ontologie WordNet est la plus complète en version anglaise. Quoi qu'il existe WOLF (WordNet Libre du Français), la version française de WordNet conçue en 2008 par l'INRIA pour la langue française, mais qui n'est pas aussi riche que l'originelle WordNet et qui est toujours en cours de développement.

Dans WordNet, une entrée est un concept représenté par un Synset (ensemble de mots ou groupes de mots, synonymes qui peuvent désigner ce concept). Les concepts reliés sémantiquement par une relation donnée à un Synset, sont représentés par une classe qui porte le nom de la relation. La relation de base entre les termes dans WordNet est la Synonymie. (WordNet est présentée dans l'annexe C).

[3]Personomie : Telle qu'elle est définie dans [Rupert, 10], la personomie Pu d'un utilisateur est Pu = (Tu,Du,Au), avec Tu l'ensemble des tags de l'utilisateur, Du l'ensemble des documents taggués par cet utilisateur et Au l'ensemble de ses annotations.

[4](Princeton University) est située dans la ville de Princeton dans l'État du New Jersey, aux États-Unis. Membre de l'Ivy League, elle a été fondée en 1746 en tant que collège du New Jersey. Elle devient l'Université de Princeton en 1896.

- **Les profondeurs des termes dans WordNet**

Comme cité ci-dessus, la composante atomique sur laquelle repose le système entier est le synset (synonym set), un groupe de mots interchangeables, dénotant un sens ou un usage particulier. La figure Fig17 illustre la définition d'un sens du mot *programming*.

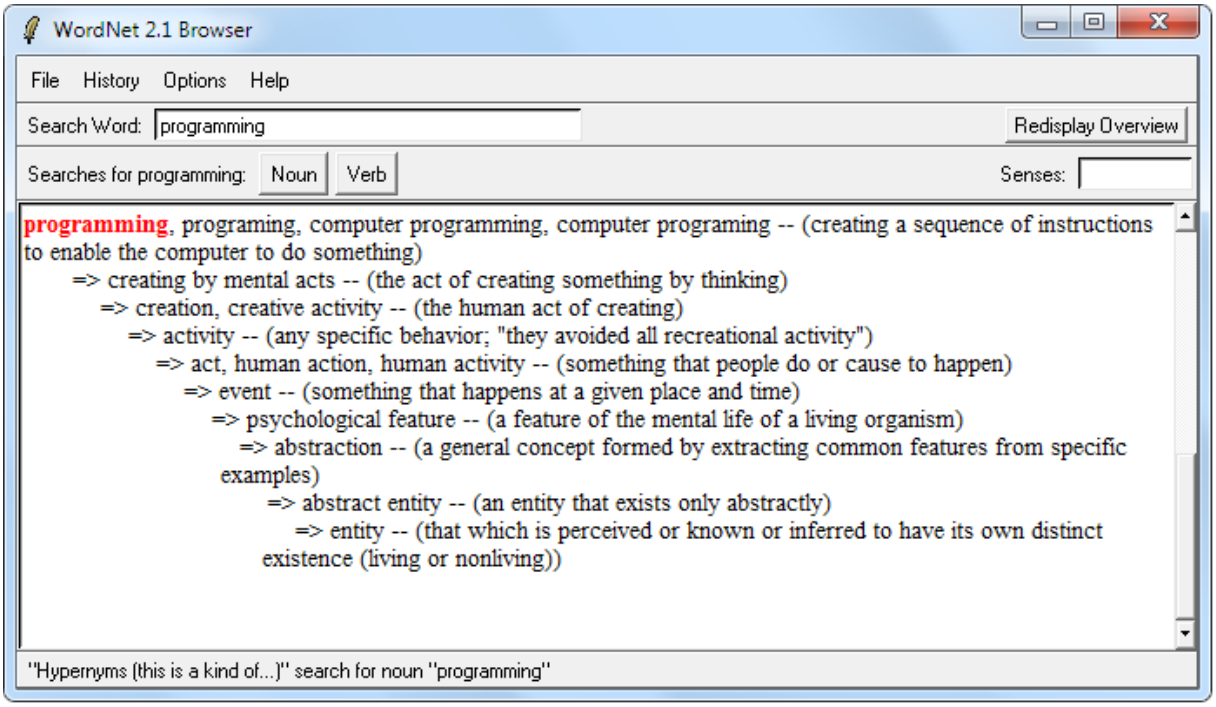

Fig 17 : *Exemple de définition de terme dans WordNet*

La profondeur d'un terme est le nombre de nœuds le séparant de la racine, dans ce cas pour *programming* la profondeur est de 9.

Un ensemble de termes (tags) et leurs profondeurs sont illustrés dans le tableau suivant.

Tag	Profondeur	Tag	Profondeur	Tag	Profondeur
Programming	**9**	reference	**7**	Tutotial	**9**
Python	**12**	Free	**5**	Web	**3**
html	**9**	Maps	**7**	Rails	**8**
Css	**9**	Data	**5**	Ajax	**9**
Design	**9**	Ruby	**9**	Gis	**8**
Geocode	**7**	Books	**9**	Video	**5**
Javascript	**8**	Google	**11**		

Tableau 05: *Exemple de calcul d'expertise utilisateur*

En appliquant notre formule de calcul de l'expertise (formule 05), l'expertise de cet utilisateur est de ***7.9.***

IV.4.2. Pondération des tags à base du profil utilisateur

Nous avons vu dans la section 8-2 du chapitre III les travaux de **[Huang, 08]** qui a introduit la notion de capacité du tag à représenter une ressource. Il s'est basé sur le fait que le premier tag cité par un utilisateur pour une ressource donnée est plus représentatif que les suivants **[Golder, 05]**. Dans notre approche, cette capacité du tag que nous appelons poids est calculée en incluant le profil utilisateur dont nous avons défini le modèle au niveau de la section précédente.

Le poids d'un tag est calculé en fonction de l'utilisateur qui l'a émis. Le même *tag* se verra donc attribuer deux poids différents si les deux utilisateurs propriétaires sont différents. D'un autre côté, pour le même utilisateur, les tags qu'il associe à une ressource devraient avoir des poids différents. Nous définissons le poids du tag en fonction du profil utilisateur représenté par ses deux dimensions centres d'intérêts et expertise. Dans l'objectif d'introduire l'aspect subjectif du tag, nous introduisons également un feedback de l'utilisateur via un rating.

Le poids d'un tag est calculé comme suit :

$$W_t^r = \sum_{i=1}^{k} \left[\left(\frac{Exp(u_i)}{dist(\overrightarrow{Interet(u_i)}, \overrightarrow{Popularity(r)})} \right)^{Conf(u_i,r)} \right]$$

Formule 06 : *Formule de pondération d'un tag à base du profil utilisateur*

Où k est le nombre d'utilisateurs ayant taggué la ressource. *dist(Interet(ui),Popularity(r))* représente le degré de rapprochement entre la ressource et les centres d'intérêts de l'utilisateur. C'est la distance entre le vecteur Interet(ui) et le vecteur de la ressource composé des tags populaires (les tags considérés comme populaires par les systèmes actuels). Cette distance est calculée avec la formule du cosinus, utilisée dans le calcul de similarité. La mesure du cosinus est donnée par la formule suivante :

$$cos(\overrightarrow{Interet(u_i)}, \overrightarrow{Popularity(r)}) = \sum_{i=1,k} \frac{wi,I}{\sqrt{\sum_{i=1,k} wi,I^2}} \frac{wi,P}{\sqrt{\sum_{i=1,k} wi,P^2}}$$

Formule 07 : *Formule de la mesure Cosinus,* ***[Gerald, 02]***

Avec $w_{i,I}$ le poids du terme d'indice i dans le vecteur $\overrightarrow{Interet(u_i)}$ et $w_{i,P}$ est le poids du terme d'indice i dans le vecteur $\overrightarrow{Popularity(u_i)}$.

La distance entre les centres d'intérêts de l'utilisateur et le contexte de la ressource est, dans notre cas, calculée comme suit :

$$dist\left(\overrightarrow{interet(u_i)},\overrightarrow{popularity(r)}\right) = 1 - \cos\left(\overrightarrow{interet(u_i)},\overrightarrow{popularity(r)}\right)$$

Formule 08 : *Formule du calcul de la distance utilisateur-ressource*

Conf(u,r) représente le degré de confiance d de l'utilisateur u dans son tag. Ceci est réalisé via un rating de 1 à 5 chaque fois que l'utilisateur taggue une ressource. Elle est calculée comme suit :

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

Formule 09 : *Formule du calcul de la confiance*

Exemple : Un exemple de mise en œuvre de cette dimension serait un rating à étoiles. Quand l'utilisateur sélectionne 5 étoiles (très confiant des tags qu'il donne à une ressource), la valeur de la confiance est maximale (1). S'il ne sélectionne aucune étoile (pas du tout confiant), la confiance prend sa valeur minimale (0).

IV.4.2.1. Discussion sur la formule de pondération des tags (formule 06)

L'expertise d'un utilisateur est calculée par rapport à tout le domaine, en la divisant par la distance de l'utilisateur avec le vecteur de la ressource, nous cherchons la part de contribution de la ressource dans l'expertise de cet utilisateur. Plus cette ressource est proche de l'utilisateur plus la distance est petite et donc le rapport est grand, ce qui fait qu'un utilisateur qui taggue une ressource proche de ses intérêts confirme son expertise et donc lui attribue un poids élevé. Alors qu'une ressource qui diverge de ses intérêts ne devrait pas avoir un grand poids sous prétexte de l'expertise de l'utilisateur dans le domaine.

Le degré de confiance de l'utilisateur dans le tag qu'il associe à la ressource est utilisé comme une sorte de régulateur du poids. Si l'utilisateur n'est pas du tout sûr de son tag, il attribue un rating de 0 et le poids calculé devient un simple calcul de popularité, alors que si

l'utilisateur attribue la note maximale, son profil est complètement utilisé dans le poids du tag. C'est donc le degré d'introduction du profil utilisateur dans le calcul du poids du tag.

Dans l'exemple suivant, nous montrons les tags de quatre utilisateurs associés à l'url http://blogs.msdn.com/jensenh/default.aspx extraite de Delicious dont le vecteur de ses tags populaires est le suivant : {design : 3, css : 2, html : 4, tools : 4}.

La figure Fig18 montre les centres d'intérêts des quatre utilisateurs calculés de la même manière que dans l'exemple de la figure Fig15. Le tableau 6, résume les tags associés par les utilisateurs à cette ressource avec leurs popularités et leurs poids calculés avec notre approche. Les distances, les expertises ainsi que les confiances des utilisateurs sont dans le tableau 7.

Fig18 : *Vecteurs d'intérêts des quatre utilisateurs*

	u1	u2	u3	u4	popularité	poids
html	x	x	x	x	4	24.31
tools	x	x	x	x	4	24.31
design		x	x	x	3	9.85
css	x		x		2	16.55
video			x		1	2.09
go		x			1	5.5
maps				x	1	2.26
presentation				x	1	2.26
programming	x				1	14.46

Tableau 06: *Liste des tags associés à la ressource*

	dist(u,r)	Exp(u)	Conf(u,r)
U1	0.28	7.9	0.8
U2	0.64	5.4	0.8
U3	0.60	3.8	0.4
U4	0.64	2.5	0.6

Tableau 07 : *Distances, Expertise et confiances des utilisateurs*

IV.4.2.1. Etude des variations de la formule de pondération

Nous étudions dans cette section les variations du poids exprimé par la formule 06 que nous avons proposée pour la pondération des tags. Le but de cette étude est de connaitre le comportement de la formule en fonction des différents paramètres.

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

En fixant les valeurs de distance à 0.1 et la confiance à 1 (utilisateur très proche du contexte de la ressource et confiant), nous varions les valeurs d'expertise de 1à 10, nous obtenons la courbe suivante.

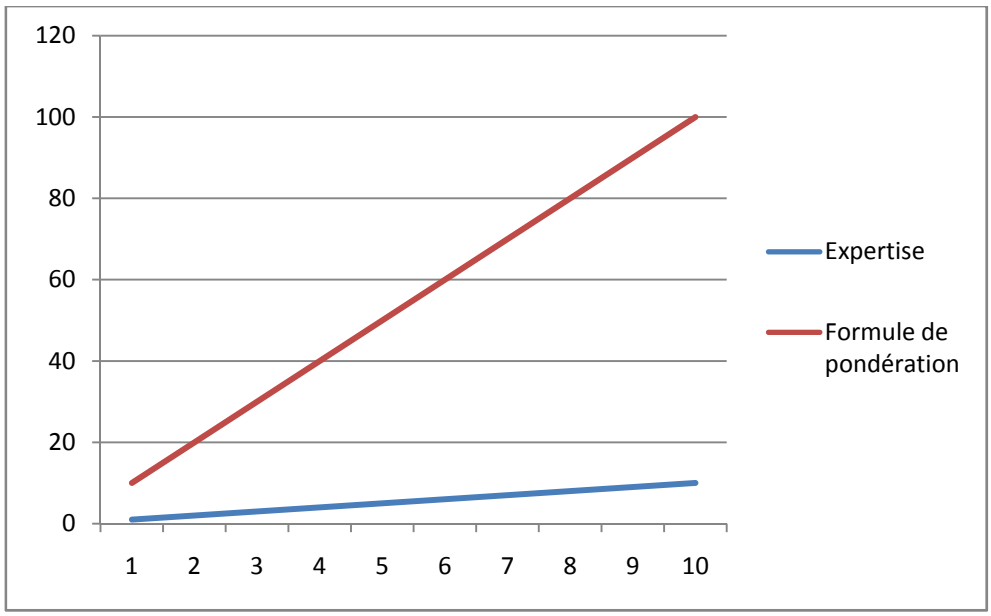

Fig19 : *Courbe illustrant les variations du poids en fonction de l'expertise*

Le paramètre expertise est très déterminant. Il est vrai que la pondération de tag dépend des trois paramètres, mais sa dépendance est très forte par rapport à ce paramètre.

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

Dans ce cas, nous fixons l'expertise à 6 et la confiance à 0.6 (utilisateur de moyenne expertise et confiance), et nous varions la distance de 0.1 à 1, nous obtenons donc la courbe suivante (Fig 20).

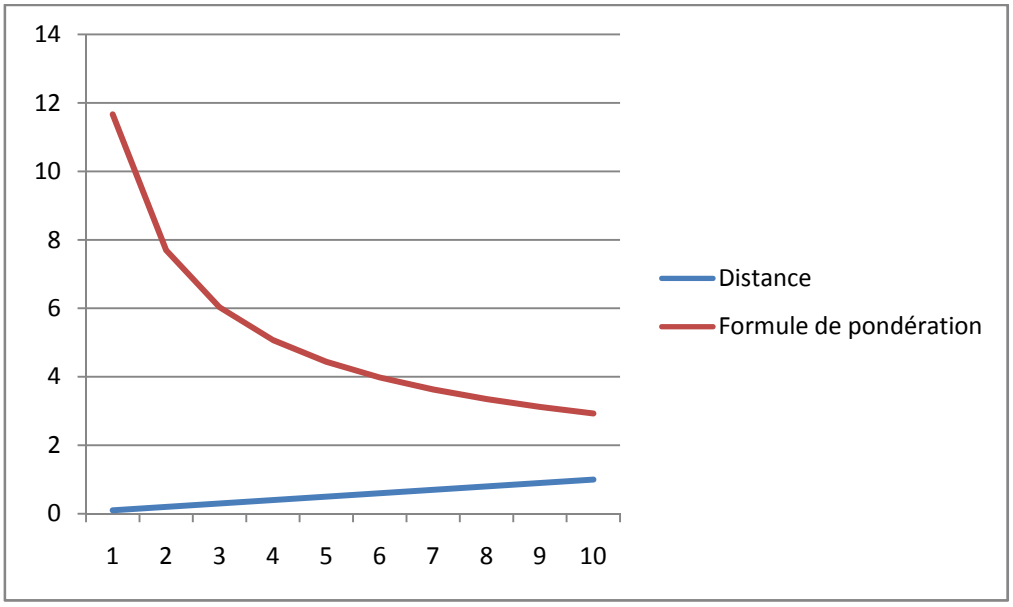

Fig20 : *Courbe illustrant les variations du poids en fonction de la distance*

La formule de pondération est inversement proportionnelle au paramètre distance (Fig 20), elle atteint son pic à la valeur minimale de la distance.

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

Dans ce cas, nous fixons le paramètre expertise à 6 et le paramètre distance à 0.5 (pour illustrer le cas d'un utilisateur moyen). La confiance prend les valeurs : 0, 0.2, 0.4, 0.6, 0.8 et 1.

La formule de pondération est proportionnelle au paramètre confiance (Fig 21), mais à un degré moins rapide le comparant au paramètre expertise.

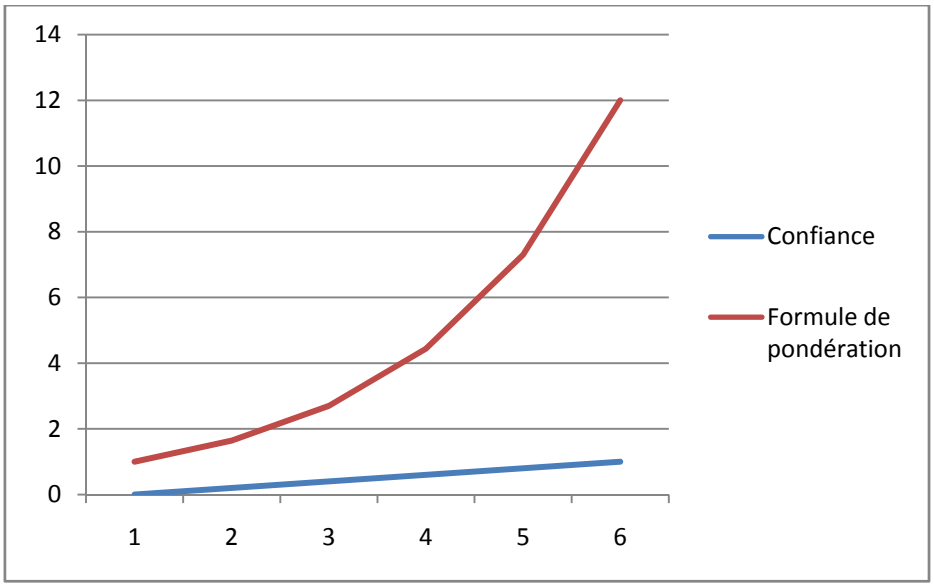

Fig21 : *Courbe illustrant les variations du poids en fonction de la confiance*

IV.4.3. Classement des tags et construction des descripteurs (filtrage)

Une ressource donnée compte un nombre considérable de tags. Après la pondération de ceux-ci avec notre approche, un classement est établi par ordre décroissant de poids w_t^r (tableau 08).

Tag	Poids (w_t^r)
html	24.31
tools	24.31
css	16.55
programming	14.46
design	9.85
go	5.5
maps	2.26
presentation	2.26
video	2.09

Tableau 08 : *Classement des tags par ordre décroissant du poids*

Le nouveau vecteur de la ressource (descripteur) est donc {html : 24.31, tools : 24.31, css : 16.55, programming : 14.46}. En comparant ce vecteur avec celui construit avec la popularité des tags, on remarque que la prise en compte du profil de l'utilisateur dans le calcul des poids favorise les tags issus des utilisateurs experts au détriment des tags plus populaires

quand ces derniers sont cités par des utilisateurs de moindre expertise. Comme est le cas pour le tag *programming* dont la popularité est plus faible que celle du tag *design* mais cité par un utilisateur plus expert. D'un autre côté, les tags ayant la même popularité se voient affectés des poids différents.

Dans le cas de cet exemple, nous avons pris les quatre premiers tags du moment que nous avons démarré d'un vecteur de ressource contenant quatre tags. Mais en réalité et comme nous l'avons déjà cité, on est confronté à un grand nombre de tags par ressource, le choix du seuil à définir comme taille du descripteur dépend du nombre total des tags utilisés dans le système. Plus le nombre de tags utilisés dans toute la Folksonomie est grand plus le seuil l'est.

Dans notre travail, et comme nous allons le voir dans la partie tests et évaluation, ce seuil est de 20 tags.

IV.5 Conclusion

La liberté du choix des tags par les utilisateurs est à l'origine de nombreux problèmes, entre autre l'attribution de tags peu représentatifs. Les tags sont classés par ordre de popularité, or, un tag populaire n'est pas forcement représentatif du contenu qu'il taggue.

Dans notre contribution, nous avons proposé une approche de pondération des tags à base du profil utilisateur dans l'objectif de créer un descripteur assez représentatif du contenu de la ressource. Nous avons donc, défini un modèle du profil utilisateur par trois dimensions : informations personnelles, centres d'intérêt et expertise ainsi qu'une approche de construction de ce dernier qui est une hybridation des deux approches naïve et par cooccurrence. Le poids du tag est calculé à base de trois facteurs, la distance entre le vecteur d'intérêts construit et le vecteur ressource composé des tags populaires, l'expertise de l'utilisateur et le facteur confiance permettant à l'utilisateur de s'évaluer par rapport à la ressource qu'il taggue.

Théoriquement, la formule proposée permet d'obtenir des poids beaucoup plus intéressants que la popularité des tags et l'introduction du profil change significativement le vecteur descriptif des ressources. Cependant, on ne peut se baser uniquement sur une étude théorique de la formule pour affirmer ses forces ou faiblesses, c'est pourquoi, nous avons implémenté un système permettant l'évaluation de cette approche et le chapitre suivant en explique les détails.

Chapitre V : Tests et Evaluations

Plus ardent qu'éclairé dans mes recherches, mais sincère en tout, même contre moi

(Jean-Jacques Rousseau, Lett. à l'archev. de Paris.)

V.1. Introduction

Pour évaluer l'efficacité de la formule de pondération proposée, nous nous sommes projetés dans un système de recherche d'information à base de tags. L'objectif étant de voir si le résultat obtenu en utilisant le nouveau poids des tags est meilleur que celui obtenu avec la popularité uniquement. Nous avons donc réalisé deux types de recherches, l'une à base de tags classés par popularité, l'autre à base de tags classés par leurs poids calculés avec notre approche. Les résultats de ces deux recherches sont comparés aux résultats d'un système de recherche d'information que nous avons également développé à cet effet.

Dans ce chapitre, nous décrivons la collection de test, la démarche d'évaluation ainsi que les résultats obtenus.

V.2. Collection de test

Actuellement, il n'existe pas de collections de test dans le domaine des travaux de recherche sur le Tagging tel que c'est le cas dans d'autres domaines comme la recherche d'information. La plupart des travaux construisent leurs propres collections, chose, à laquelle nous avons été confrontés.

Nous avons effectué les tests sur une collection de 149 URLs récupérées de Del-icio-us, tagguées par 6 utilisateurs de profils complètement différents en utilisant 215 tags. 565 actions de Tagging ont été effectuées par les 6 utilisateurs avec en moyenne 43,66 URLs tagguées par utilisateur.

V.3. Démarche d'évaluation

La procédure d'évaluation suivie passe par trois phases principales, la phase de préparation de la collection, la phase d'utilisation du SRI la phase de comparaison de résultats, nous illustrons globalement ces phases dans le schéma ci-dessous et nous détaillons par la suite chacune de ces étape.

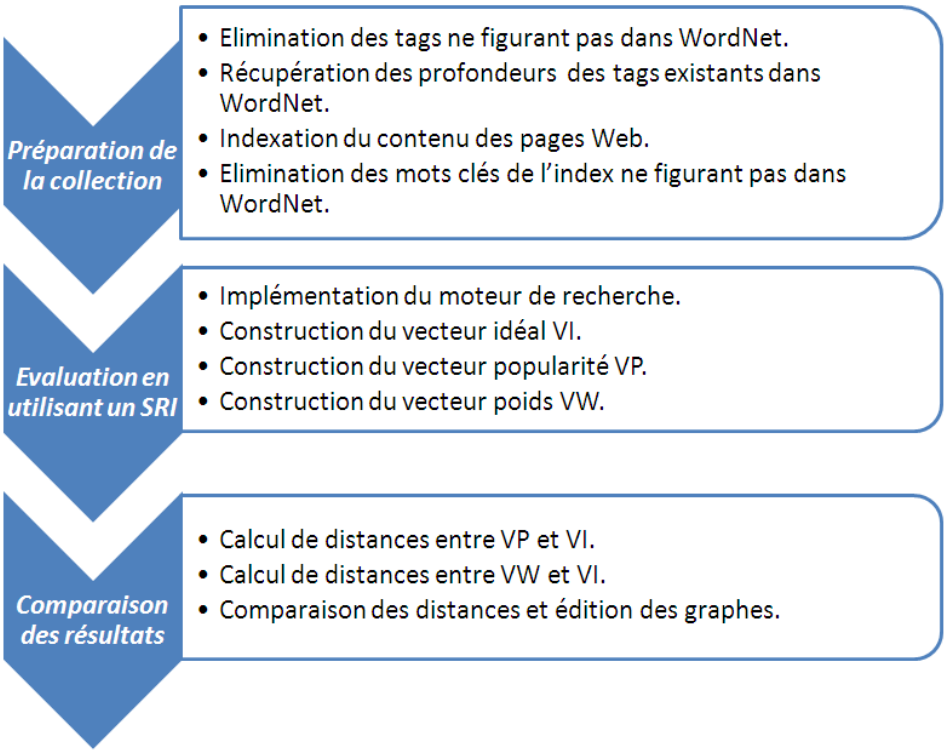

Fig 22 : *Processus d'évaluation de l'approche*

V.3.1. Phase de préparation de la collection

Cette phase consiste à préparer la collection des URLs pour effectuer les tests. Il s'agit entre autres d'effectuer les actions suivantes :

V.3.1.1. Elimination des tags ne figurant pas dans WordNet

Les tags utilisés par les utilisateurs ne figurant pas dans WordNet sont éliminés, vu que ces derniers ne sont pas considérés comme porteurs d'information et que leur profondeur ne peut être déterminée.

V.3.1.2. Récupération des profondeurs des tags à partir de WordNet

La profondeur d'un terme dans Wordnet est le nombre de nœuds le séparant de la racine, ainsi, pour chaque tag de la collection, nous avons déduit sa profondeur en choisissant le sens adéquat selon le contexte de la page web.

Le diagramme d'activité suivant illustre les deux étapes précédentes :

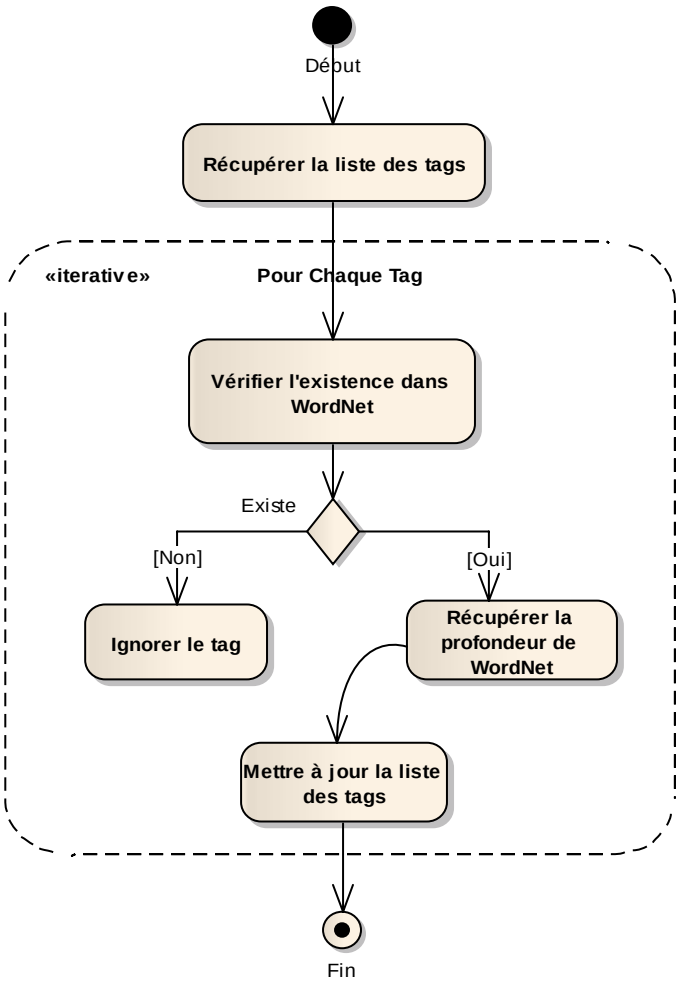

Fig 23 : *Diagramme d'activité du processus de préparation des tags*

V.3.1.3.Indexation du contenu textuel des pages web

Les pages web utilisées dans la collection ont toutes été téléchargées et leur contenu indexé. Nous avons développé à cet effet un moteur d'indexation basé sur la technique du fichier inverse.

Le processus d'indexation consiste d'abord à extraire les termes du document n'apparaissant pas dans un anti-dictionnaire (*Stop-Liste*[5]) et dans une certaine limite de fréquences **[Dahak, 06]**. Les termes sélectionnés sont ensuite lemmatisés, ce qui permet une certaine normalisation des termes dans l'index **[Paradis, 96]**.

L'index résultant est sauvegardé pour être utilisé dans la recherche. La pondération des termes est réalisée avec la formule du Tf/Idf. (Pour plus de détails sur l'indexation, voir l'annexe D).

[5] Stop-Liste : Liste des mots fréquemment utilisés dans une langue donnée. Donc insignifiants en terme de description du contenu d'un document. Ex : la, le, de, dans, en, je…

Le diagramme d'activité suivant montre le processus d'indexation implémentée dans notre système :

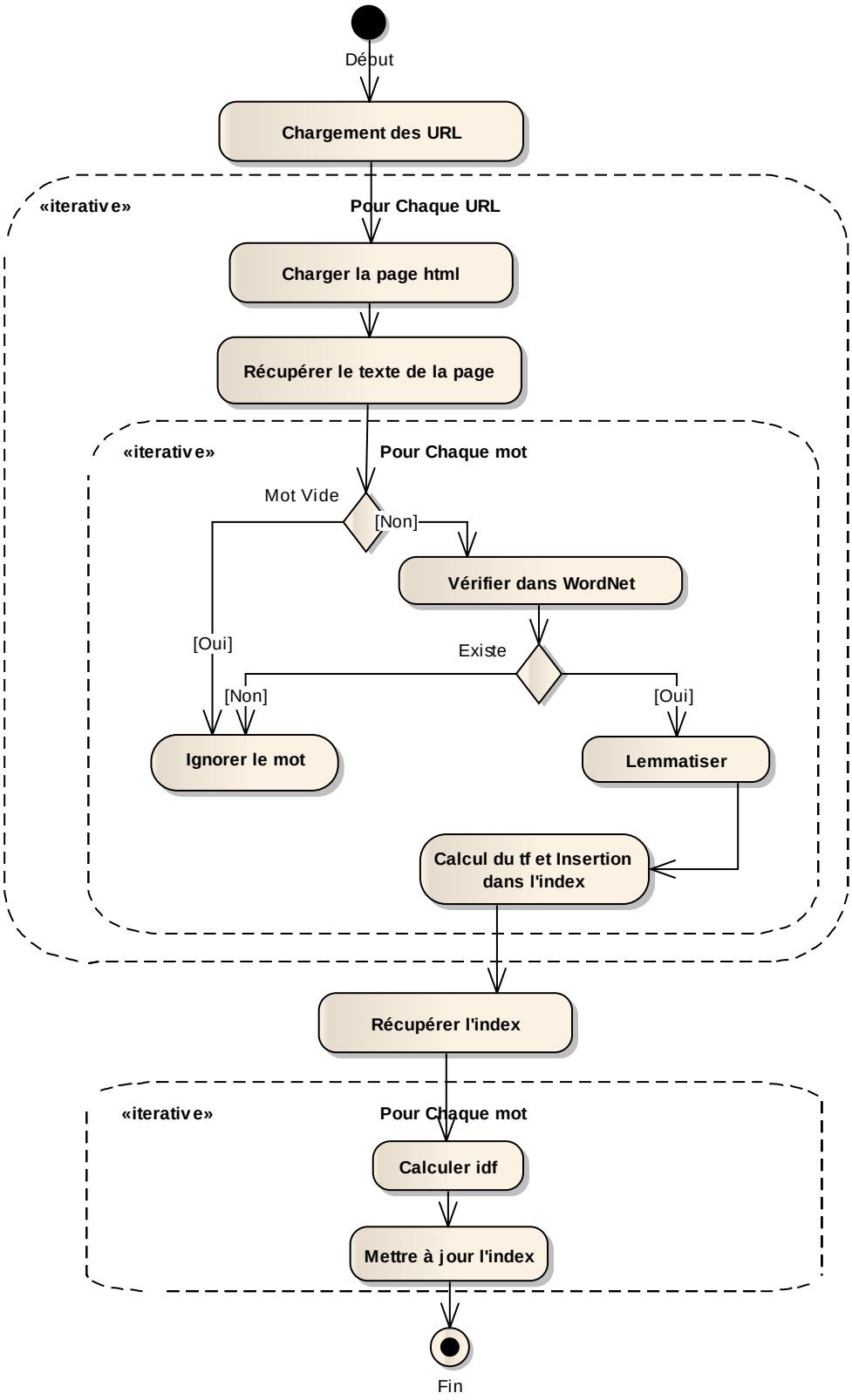

Fig 24 : *Diagramme d'activité du processus d'indexation*

V.3.1.3. Elimination des mots-clés de l'index ne figurant pas dans WordNet

Une fois l'indexation terminée, le fichier inverse obtenu est purifié en enlevant tous les mots ne figurant pas dans WordNet. Cette action permet d'avoir un équilibre entre la représentation des pages web avec les tags des utilisateurs et l'index textuel.

V.3.2. Evaluation en utilisant un système de recherche d'information (SRI)

Cette partie de notre système d'évaluation permet d'effectuer une recherche sur l'index et récupérer une liste d'URLs ordonnée par ordre de pertinence répondant à une requête utilisateur. Le modèle de recherche d'information que nous avons adopté est le modèle vectoriel.

V.3.2.1. Implémentation du moteur de recherche

Le principe de base du modèle vectoriel (Annexe D) est d'utiliser une représentation géométrique pour classer les documents par ordre de pertinence par rapport à une requête. Les documents et les requêtes sont représentés par des vecteurs dans un espace vectoriel de n dimensions, chaque dimension est une caractéristique du document.

Le processus de recherche est représenté dans le diagramme d'activité suivant :

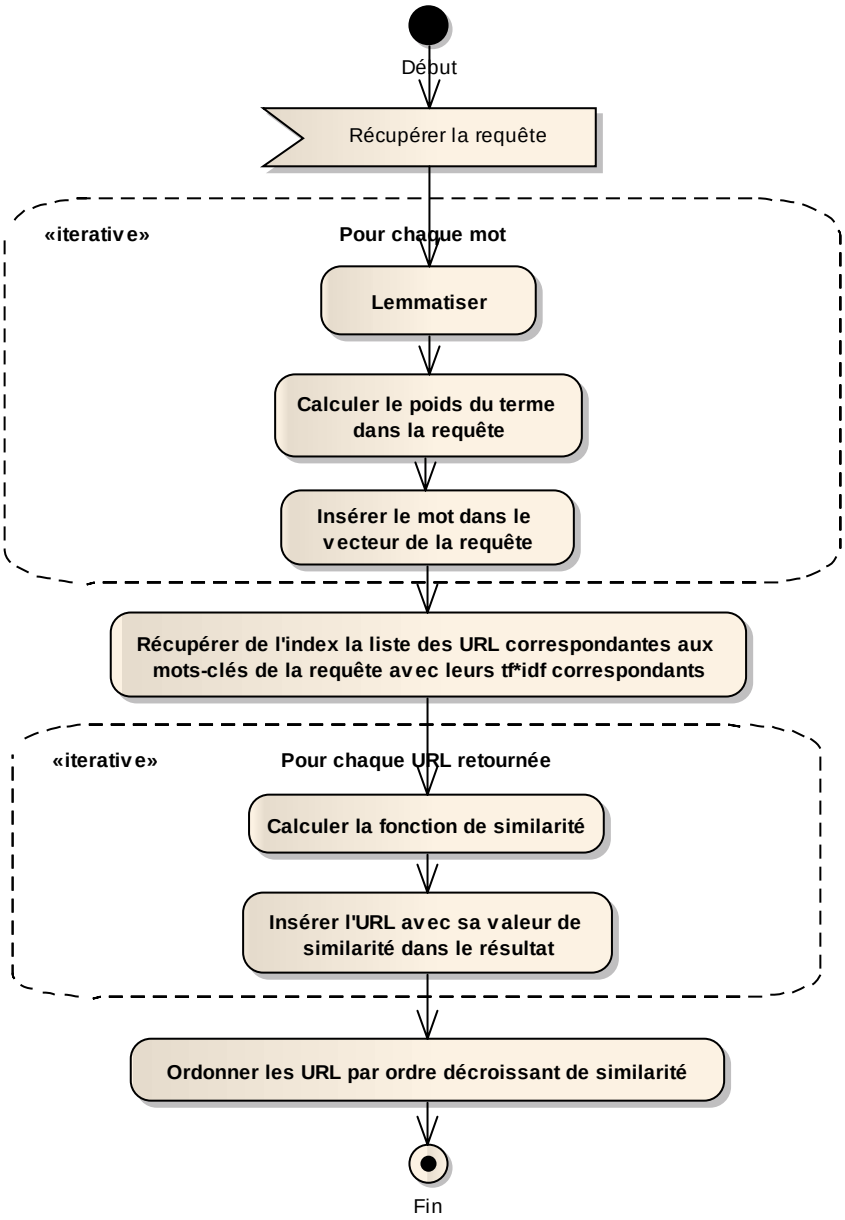

Fig 25 : *Diagramme d'activité du processus de recherche*

V.3.2.2. Construction du vecteur idéal

Pour chacun des 50 premiers tags les plus populaires, nous avons construit un vecteur idéal comme suit : Le tag en question est considéré comme la requête pour le moteur de recherche qui effectue une recherche dans l'index et récupère la liste des Urls correspondantes au tag introduit (Fig 26). La liste est ordonnée par ordre décroissant de pertinence. La figure ci-dessous représente le résultat de la recherche du tag « python » à partir de notre moteur de recherche.

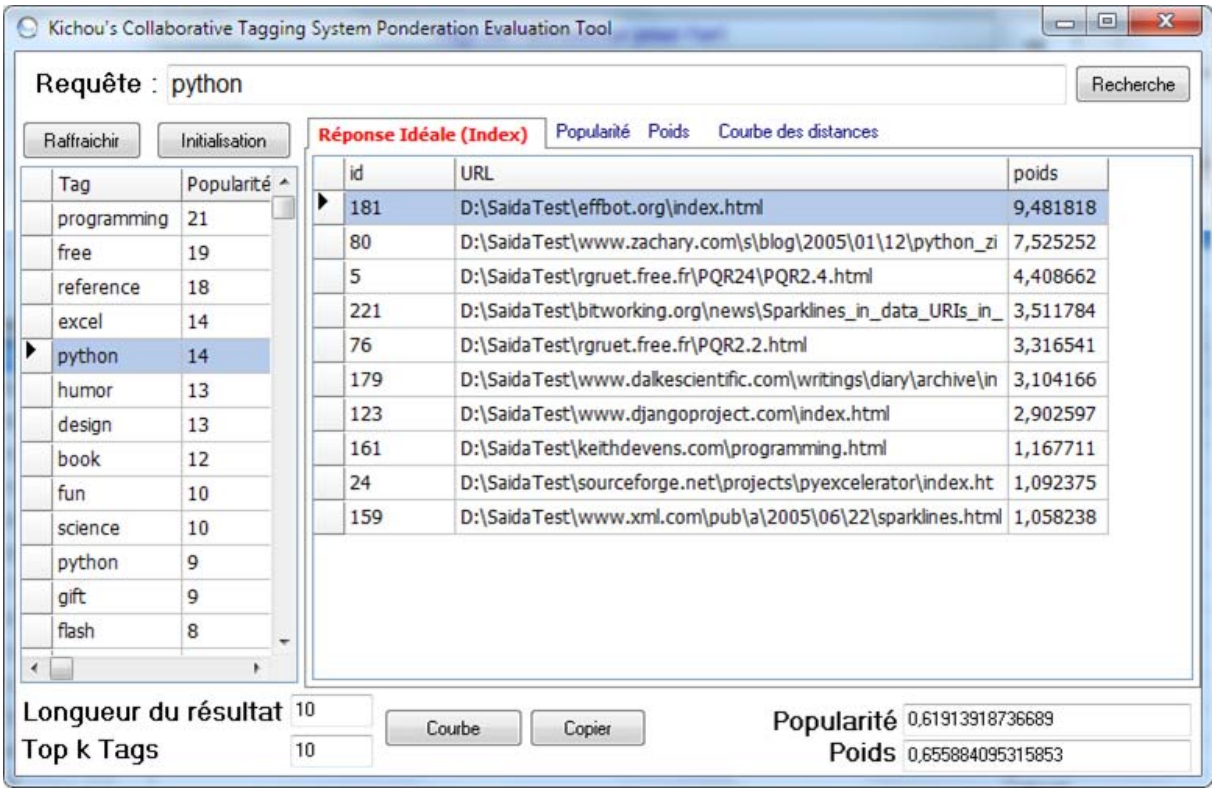

Fig 26 : *Construction du vecteur idéal **(VI)***

V.3.2.3. Construction du vecteur popularité

Pour un tag donné, le système cherche dans la base de données les URLs qui ont été tagguées avec le tag en question et récupère sa popularité. Les URLs résultantes sont ensuite classées par ordre décroissant de popularité. Dans ce cas la popularité est le nombre de fois que le tag a été utilisé pour tagguer l'URL. La figure suivante représente le résultat d'une recherche à base de popularité pour le même tag que l'exemple précédent : « python ». Le résultat de cette recherche construit le vecteur popularité (**VP**).

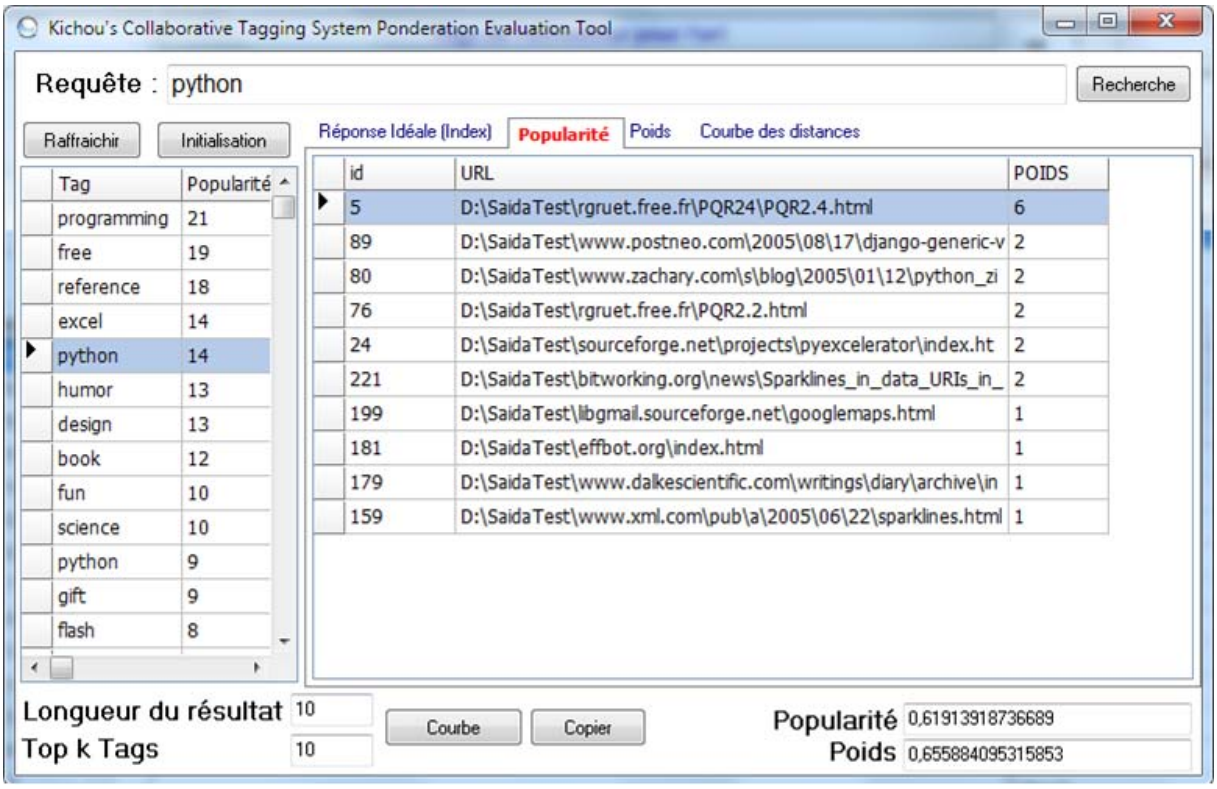

Fig 27 : *Construction du vecteur popularité* ***(VP)***

V.3.2.4. Construction du vecteur poids

De la même manière que pour la popularité, nous construisons le vecteur poids mais cette fois-ci les URLS résultantes ne sont pas classées par ordre de popularité mais par rapport aux poids calculés avec notre formule. Le résultat de la recherche forme le vecteur Poids (**VW**). La figure ci-dessous représente le résultat de la recherche du tag « python » à base du poids des tags.

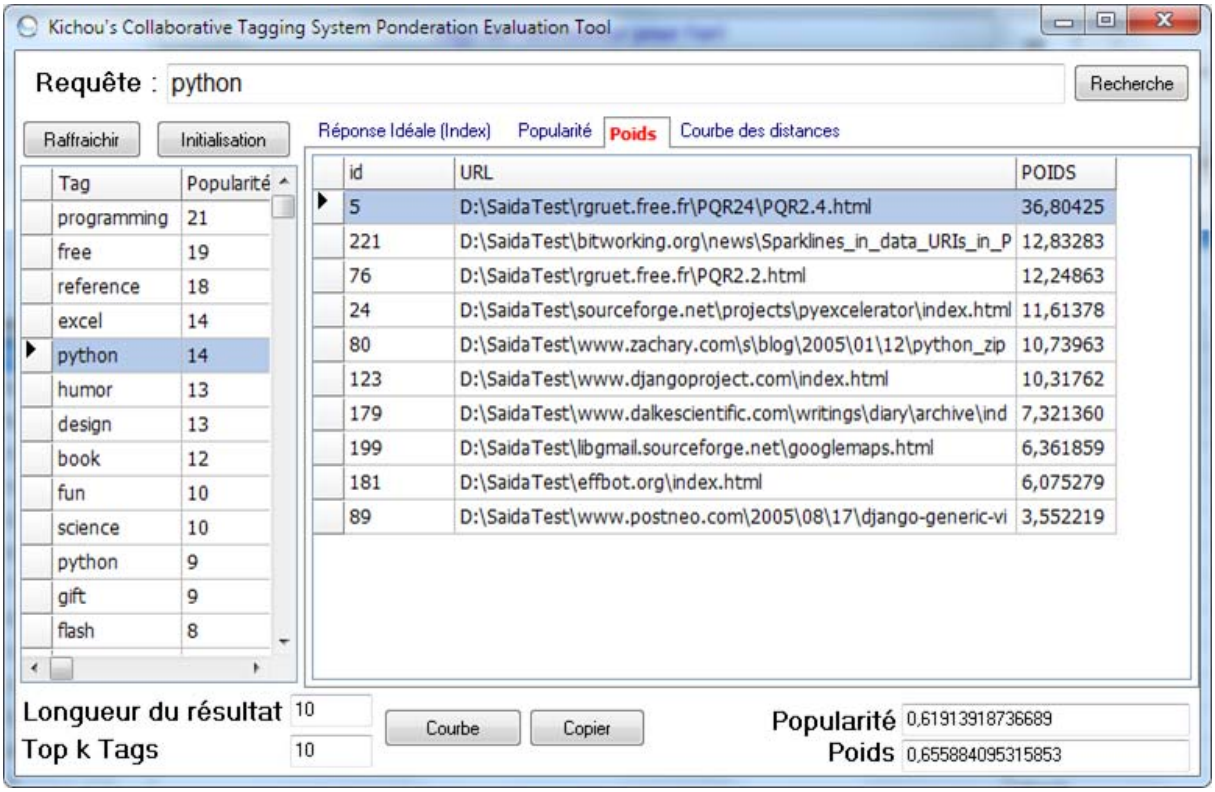

Fig 28 : *Construction du vecteur poids (**VW**)*

V.3.3. Comparaison des résultats

L'objectif de la comparaison étant de voir lequel des deux vecteur obtenus popularité **VP** ou poids **VW** est le plus proche du vecteur idéal **VI** obtenu à partir de l'index.

Nous considérons le vecteur obtenu à partir de l'index comme étant idéal parce que ce dernier est obtenu à partir des mots-clés constituant le contenu de la page web. Il représente donc le contenu de la page mieux que les tags donnés par les utilisateurs qui ne sont pas forcement représentatifs. Cependant, un tag utilisé par plusieurs utilisateurs a de grandes chances d'appartenir à l'index de cette dernière, dans quel cas, on cherche à savoir le quel des deux rangs, la popularité ou notre poids, fait que le classement de l'URL dans la liste des résultats soit le plus proche possible de sont classement dans l'index.

Nous avons donc construit les trois vecteurs pour chacun des 50 premiers tags et nous avons calculé la distance du cosinus entre le vecteur idéal et le vecteur popularité, ensuite entre le vecteur idéal et le vecteur poids. Les distances obtenues ont été représentée dans un graphe afin d'en illustrer la comparaison. Rappelons que plus la distance est grande plus les vecteurs comparés sont proches. La figure ci-dessous représente le résultat de la comparaison des deux vecteurs dans notre évaluation.

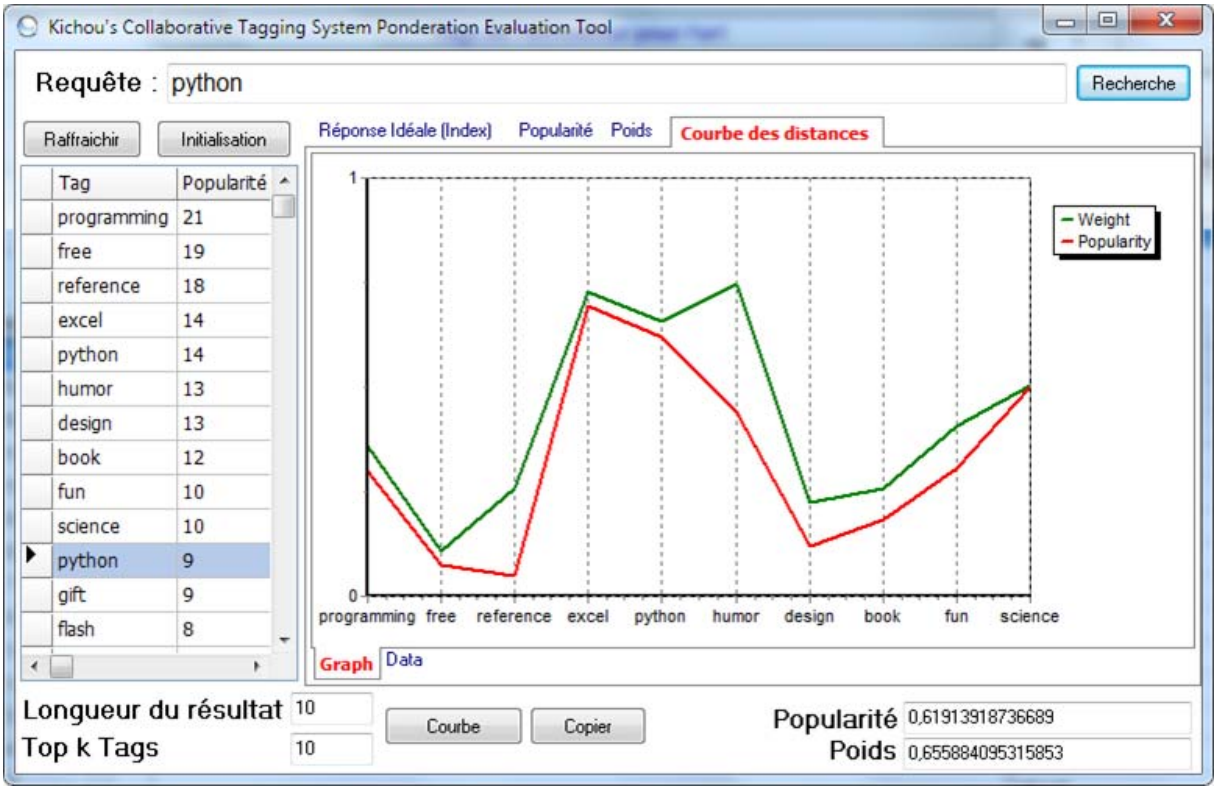

Fig 29 : *Courbe de comparaison des vecteurs*

V.4. Architecture du système d'évaluation

Le système développé permet à la fois l'action du Tagging (Ajout et modification de tag) et l'action d'évaluation qui consiste en les trois phases présentées précédemment. La figure ci-dessous, illustre l'architecture de notre système d'évaluation.

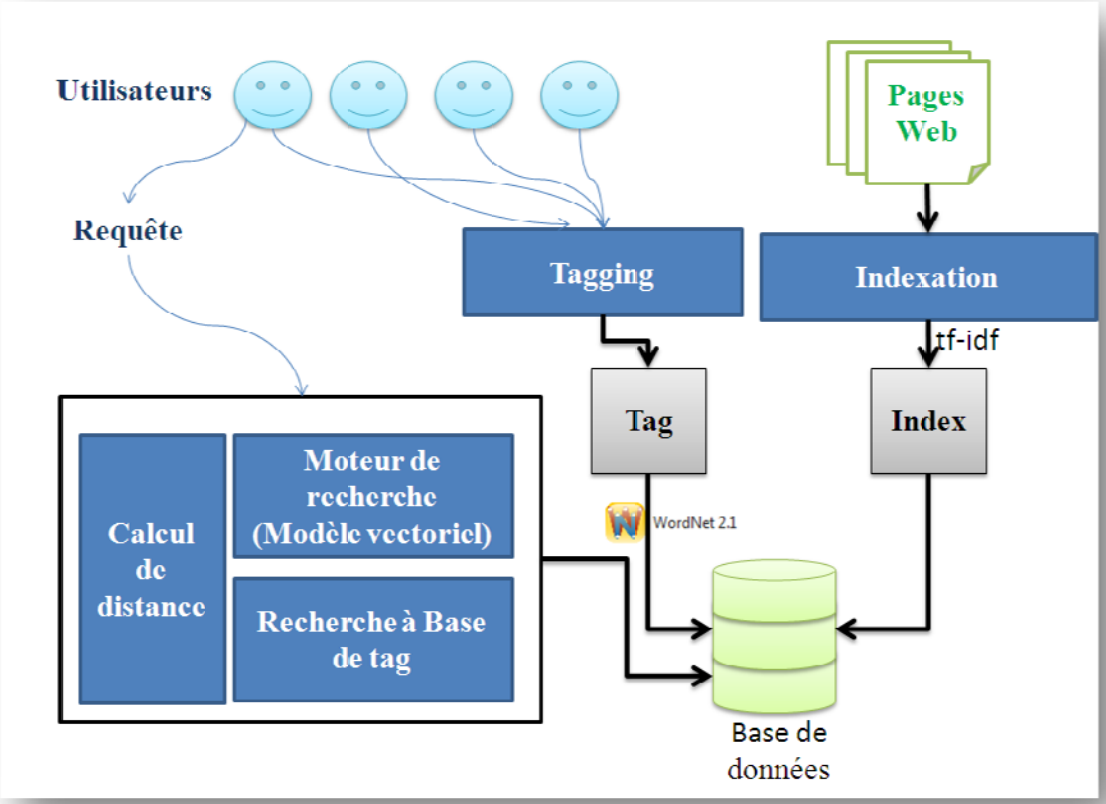

Fig 30 : *Architecture du système d'évaluation*

Le système permet, entre autres, les fonctionnalités suivantes:

- ***Ajout d'un utilisateur :*** Consiste à inscrire un utilisateur en introduisant les informations suivantes : nom, prénom, login et mot de passe.
- ***Ajouter et tagguer une ressource :*** Permet de partager une ressource donnée et lui associer un ensemble de tags.
- ***Opérations du Tagging :*** Permet d'associer des tags aux ressources, les contenus de celles-ci peuvent être affichés par les utilisateurs.
- ***Création des vecteurs Intérêt :*** C'est une mise en œuvre de l'approche hybride proposée, pour construire la dimension centres d'intérêts associée à chaque utilisateur.
- ***Indexation des pages web :*** Permet la création des index des pages web.
- ***Effectuer les recherches et calculer les distances :*** C'est l'étape de construction des vecteurs VI, VP et VW et comparaison de ceux-ci. Cette étape est citée précédemment dans ce chapitre.

Le diagramme des cas d'utilisation suivant montre clairement ces fonctionnalités :

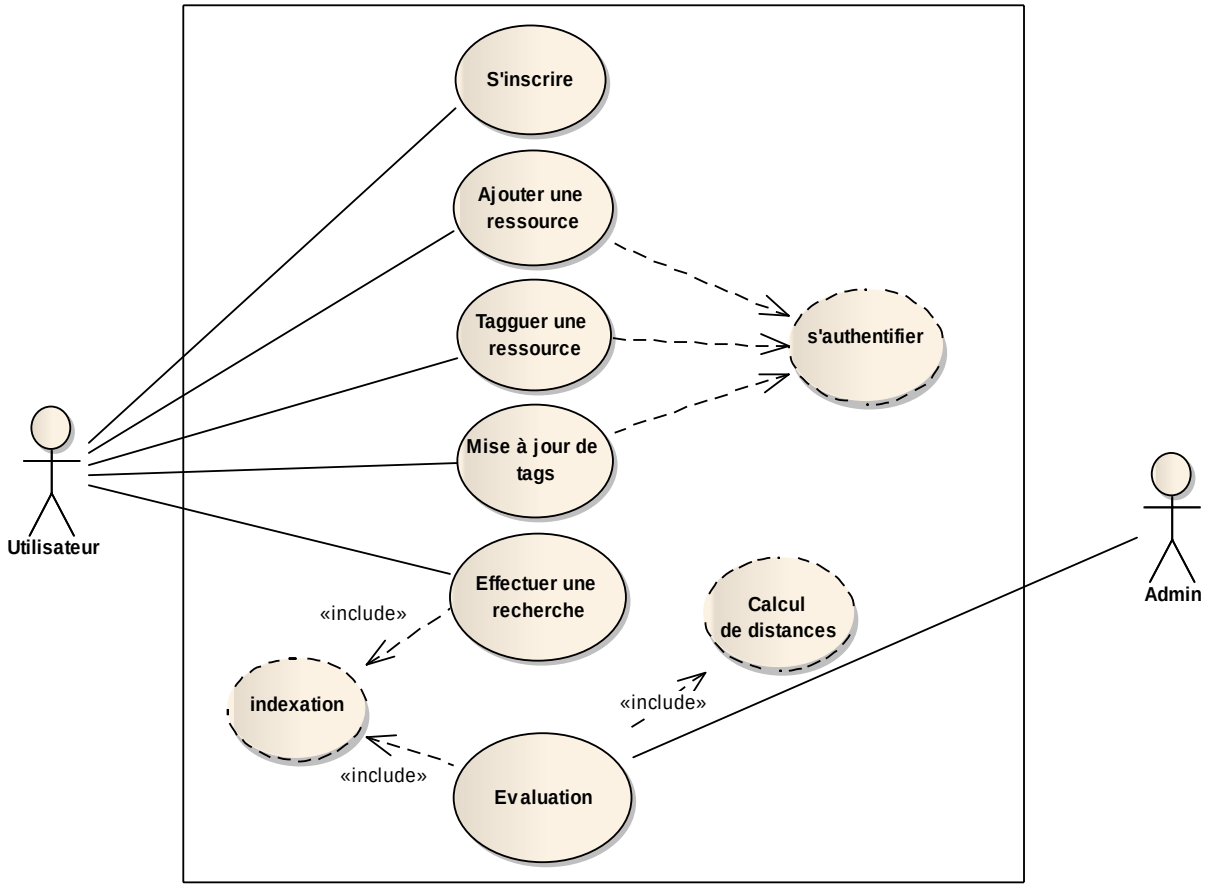

Fig 31 : *Diagramme des cas d'utilisation*

La structure de la base de données est représentée dans le diagramme des classes de la figure (Fig 32) ci-dessous. Les ressources sont contenues dans la classe URL. L'association « tagging » concrétise l'action du tagging (association d'un tag à une ressource par un utilisateur) et la classe association « tagging_action » modélise la confiance qu'exprime l'utilisateur vis-à-vis de ses tags attribués à l'url en question. La propriété distance de cette classe représente la distance calculée entre le vecteur de l'URL et le vecteur intérêt de l'utilisateur, cette information est utile dans la formule de calcul du poids. La classe « word_index » représente l'index proprement dit du contenu textuel des pages web. c'est en fait une classe qui représente le fichier inverse alors que la classe « tag_index » comptabilise pour chaque couple « tag,url » la popularité et le poids du tag selon notre formule.

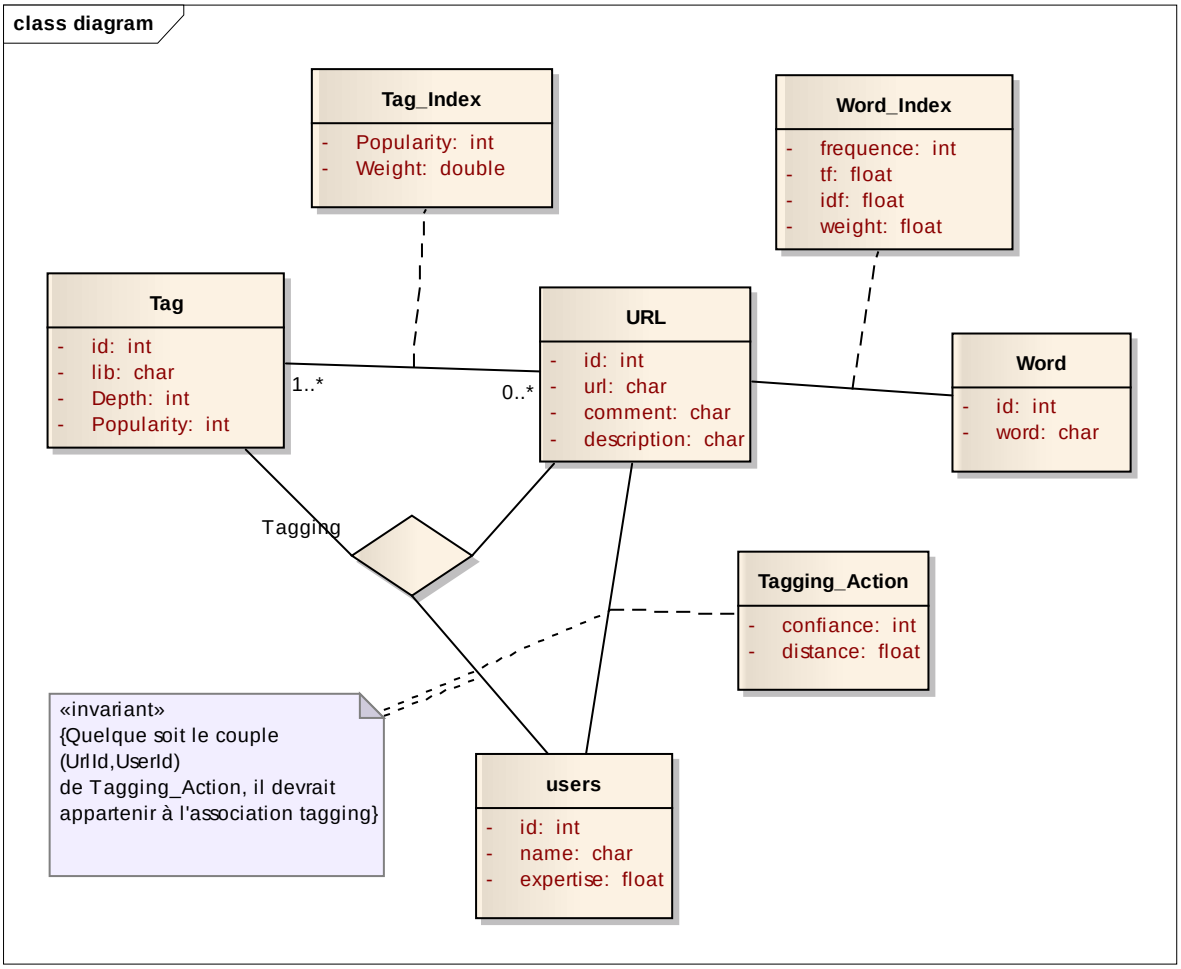

Fig 32 : *Diagramme des classes*

V.5. Résultats et discussion

Plusieurs tests ont été effectués sur la collection en paramétrant le nombre de tags sur lesquels effectuer la recherche. A partir des différents résultats obtenus, nous nous somme rendu compte que les résultats des deux recherches (par popularité et par poids) convergent en dépassant un certain nombre de tags (généralement 10% du nombre total de tags). Ceci s'explique par le fait que les tags classés en bas de liste ont une popularité faible et par conséquent un poids aussi faible vu que le nombre d'utilisateurs les ayants utilisé diminue avec la popularité. La courbe ci-dessous (Fig 33) est obtenue avec 10% des tags de la collection. Les distances sont représentées dans le tableau 9.

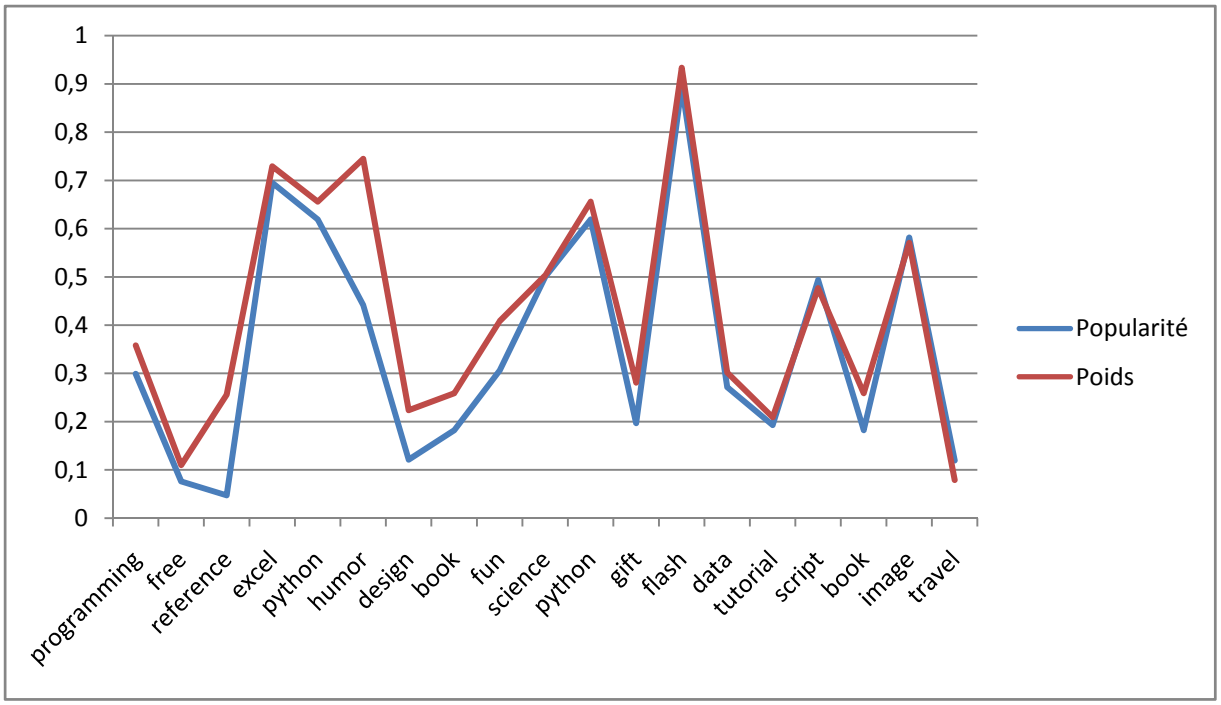

Fig33 : *Comparaison entre la recherche à base de la popularité et du nouveau poids des tags*

Tag	Dist(VI,VP)	Dist(VI,VW)
programming	0,2990743	0,3580088
free	0,07597372	0,10993814
reference	0,04714045	0,25592397
excel	0,69508351	0,72904193
python	0,61913919	0,6558841
humor	0,44151079	0,74454538
design	0,12121212	0,22361631
book	0,18226448	0,25893115
fun	0,30618622	0,40824829
science	0,5	0,50300123
python	0,61913919	0,6558841
gift	0,19676759	0,28070175
flash	0,89642146	0,93334827
data	0,27124449	0,30162774
tutorial	0,19287919	0,20889987
script	0,49382916	0,47673129
book	0,18226448	0,25893115
image	0,58157116	0,57057465

travel	0,11952286	0,07902167

Tableau 09 : *Distances (cosinus) entre les vecteur VP, VW et VI*

Nous remarquons que les résultats obtenus avec notre formule sont significativement meilleurs que ceux obtenus avec la popularité uniquement. Cependant la qualité des utilisateurs est un facteur non négligeable et qui, dans un autre contexte, pourrait dégrader les résultats.

V.6. Conclusion

Pour évaluer notre approche nous nous sommes projetés dans un système de recherche d'information à base de tags. L'objectif étant de voir si le résultat obtenu en utilisant le nouveau poids des tags est meilleur que celui obtenu avec la popularité uniquement, nous avons ainsi développé un système de recherche d'information en implémentant le modèle vectoriel. La recherche à base de l'index est donc prise comme résultat idéal par rapport auquel nous avons comparé la recherche à base de popularité et à base du nouveau poids. Les évaluations que nous avons effectuées montrent une nette amélioration des résultats de recherche en utilisant le nouveau poids.

Cependant, il est à noter que la qualité des utilisateurs joue un rôle non négligeable et que dans certains contextes, les résultats pourraient se détériorer. Ceci est le côté négatif de la subjectivité de la formule représentée par la confiance donnée par l'utilisateur.

Conclusion Générale

Chercher n 'est pas une chose et trouver une autre,

mais le gain de la recherche, c 'est la recherche même.

Saint Grégoire de nysse (Hometies sur l 'Ecclésiaste)

Synthèse

Le travail que nous avons présenté dans ce mémoire s'inscrit dans le domaine du web 2.0, et plus particulièrement du Tagging collaboratif. Ce nouveau paradigme a vu le jour ces dernières années, et a vite gagné de la popularité auprès des utilisateurs qui ont trouvé en lui un moyen d'organiser leurs ressources en ligne de manière collective.

Nous avons donc présentés les différentes notions liées au Tagging, commençant par l'annotation, ses définitions, modèles et structures. Nous avons par la suite étudié le Tagging collaboratif, ses modèles, ses propriétés et nous avons cité des exemples de systèmes de Tagging collaboratif existants et recensé quelques limites. Nous avons établi un petit comparatif entre la notion d'annotation et celle du Tagging, tant confondus dans la littérature.

Parmi les problèmes rencontrés dans les systèmes du Tagging collaboratif, la définition des tags les plus appropriés pour une ressource donnée, nous avons proposé d'intégrer le profil utilisateur pour sélectionner les meilleurs tags. La notion du profil utilisateur a été donc présentée, sa définition, modélisation, dimensions et ses différentes exploitations dans le Tagging collaboratif sont également évoquées.

Résumé de la contribution

Dans le cadre de ce travail, nous avons proposé une approche de filtrage de tags à base du profil utilisateur, pour cela nous avons d'abord défini un modèle utilisateur susceptible de contenir les informations que nous estimons essentielles pour mesurer son aptitude quant à la proposition de tags pour une ressource donnée. Trois dimensions sont définies, la dimension personnelle, les centres d'intérêt et l'expertise. Nous avons proposé une hybridation de deux approches de construction du profil utilisateur à base des opérations du Tagging qu'il effectue. Et nous avons proposé également une formule pour le calcul de l'expertise de l'utilisateur. Et enfin, une nouvelle formule de pondération de tags basée sur trois facteurs a été proposée : le premier facteur est l'expertise de l'utilisateur, le deuxième est la distance entre ses centres d'intérêt et le contexte de la ressource, le dernier est sa confiance quant aux tags qu'il associe à cette ressource.

L'évaluation de notre approche sur une collection de 169 pages web extraites de Del-icio-us, projetée sur un système de recherche d'informations a montré des résultats satisfaisants.

Perspectives

Les perspectives envisageables pour notre travail sont :

- Evaluer l'approche sur une collection plus importante, que ce soit en nombre de ressources, de tags ou d'utilisateurs ;
- Extension de l'approche de manière à résoudre les problèmes de variations d'écriture, en effet deux tags considérés différents par le système, alors qu'en réalité est un même tag, est à l'origine de nombreux problèmes entre autres le silence (non retour d'informations pertinentes) pour un besoin donné en information ;
- L'utilisation d'une ontologie de domaine pour une meilleure définition de la notion d'expertise, en effet dans notre mise en œuvre nous avons utilisé WordNet pour récupérer les profondeurs de tags. L'application de l'approche sur une collection de ressources d'un domaine donné et l'utilisation d'une ontologie de ce domaine pourrait avoir des résultats plus intéressants ;
- Un des problèmes que nous avons rencontré dans la réalisation de notre approche, est le choix du terme (le tag) dans l'ontologie utilisée vu qu'un tag donné peut avoir plusieurs sens, et donc plusieurs niveaux d'hiérarchies, ce qui donne des profondeurs différentes. Une solution serait de mettre en œuvre un Tagging guidé par une ontologie afin de définir au préalable la signification du tag utilisé.

Bibliographie

[Abrouk,10] L. Abrouk, D. Amblard and D. Leprovost. *Découverte de communautés par analyse des usages.* Workshop The Web Social - EGC 2010.

[Adler, 72] Adler.M, van Doren.C: *How to read a book*. Simon and Schuster, New York, 1972.

[Amardeilh, 07] Amardeilh .F *: Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle.* Thèse de doctorat, Université Paris X,2007.

[Amato, 99] Amato.G, U. Straccia, *User Profile Modeling and Applications to Digital Libraries*, In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France, 1999.

[Amer-Yahia, 08] S.Amer-Yahia, A.Galland, J. Stoyanovich, Cong Yu: *From del.icio.us to x.qui.site: recommendations in social tagging sites.* SIGMOD'08, June 9–12, 2008, Vancouver, BC, Canada.

[Atilf, 92] Atilf: *Le trésor de la langue française*. Paris, France. Unité mixte de recherche ATILF (Analyse et Traitement Informatique de la Langue Française). URL : http://atilf.atilf.fr/ .

[Attar, 04] P.Attar: *RDF, Ressource Description Framework. http://www.tireme.fr/glossaire/SPEC-RDF.pdf*

[Azouaou, 06] F. Azouaou : *Modèles et outils d'annotations pour une mémoire personnelle de l'enseignant* » Thèse de doctorat en Informatique, Université Joseph Fourier – Grenoble I, soutenue le 19 Octobre 2006.

[Azouaou, 05] Azouaou.F, chen.W, Desmoulins.C: *Semantic annotation tools for learning materiel: specification and categorization*. CAiSE'05(The 17th Conference on Advanced Information Systems Engineering), Porto, Portugal, 2005.

[Baldonado, 00] Baldonado.M, Cousins.S, Gwizdka.J, Paepcke.A: *Notable: At the intersection of annotations and handheld technologies*. Proceedings of HUC conference, LNCS 1927, Springer Verlag, Berlin, 2000.

[Bao, 07] S.Bao, X. Wu, B. Fe, G.Xue, Z.Su, Y.Yu: *Optimizing Web Search Using Social Annotations.* International World Wide Web Conference

Committee (IW3C2), Banff,Alberta, Canada, 2007.

[Baziz, 05] M.Baziz: *Indexation conceptuelle guidée par ontology pour la recherché d'information*. Thèse de doctorat en informatique, université Paul Sabatier, soutenue le 14 décembre 2005.

[Benna, 08] A.Benna: *Annotation et interrogation de données non structures: Application aux services web*. Thèse de magister, Université Abderrahmane Mira de Bejaia, 2008.

[Bischoff, 08] K. Bischoff, Claudiu S. Firan, W. Nejdl, R. Paiu : *Can All Tags be Used for Search?* Proceeding of the 17th ACM conference on Information and knowledge Management, 2008.

[Boufaida, 08] Z.Boufaida: *le web sémantique.* Laboratoire Lire Université Mentouri de Constantine.

[Boulkrinet, 07] S.Boulkrinet: *Modélisation hybride du profil utilisateur pour un système de filtrage d'informations sur le web.* Thèse de magister en informatique, Ecole nationale supérieure d'informatique, 2007.

[Bouzeghoub, 05] Bouzeghoub.M, D.Kostadinov: *Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils.* In Proceedings of Actes de la Conférence francophone en Recherche d'Information et Applications CORIA'2005. pp.201~218.

[Bringay, 03] Bringay.S, Barry.C, Charlet.J: *Les documents et les annotations du dossier patient hospitalier*. Information-interaction-intelligent. 2003.

[Bringay, 06] Bringay.S : *Les annotations pour supporter la collaboration dans le dossier patient électronique.* Thèse de doctorat en informatique, Université de Picardie Jules Verne – Amiens, soutenue le 04 septembre 2006.

[Broudoux, 06] Broudoux.E : *Folksonomie et indexation collaborative, rôle des réseaux sociaux dans la fabrique de l'information*. Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland, May, 2006.

[Brusilovsky, 01] P.Brusilovsky: *Adaptive Hypermedia , Adaptive Hypertext and user Modelling and user adapted interaction*, V11 p 87-110, 2001

[Cattuto, 07] Cattuto.C, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. *Network properties of folksonomies.*

AI Communications Journal, Special Issue on"Network Analysis in Natural Sciences and Engineering", 2007.

[Carmagnola, 08] Carmagnola.F, F. Cena, L. Console, O. Cortassa, Cristina Gena, Anna Goy, Ilaria Torre : *Tag-based User Modeling for Social Multi-Device Adaptive Guides*. Special issue on Personalizing Cultural Heritage Exploration, 2008.

[Cayzer, 09] Cayzer.S, E. Michlmayr : *Adaptive user profiles : chapitre de livre Collaborative and social Information Retrieval and Access*, ISBN-13: 9781605663067, 2009.

[Ciravegna, 02] Ciravegna.F, Dingli.A, Petrelli.D, Wilks.Y: *Timely and non intrusive active document annotation via adaptive information extraction*. In proceedings of semantic authoring, annotation and knowledge markup workshop. In ECAI, Lyon, France, 2002.

[Dahak, 06] F.Dahak: *Indexation des documents semi-structurés*. Thèse de magister, Ecole supérieure d'informatique, 2006.

[Damas, 02] Damas.L, Mille.A, Versace.R : *Prendre en compte les comportements cognitifs des apprenants dans la conception des systèmes d'assistance à l'apprentissage humain*. TICE, Lyon, France, 2002.

[Demontils, 02] Demontils.E, Jacquin.C : *Annotations sur le Web : notes de lecture*. AS CNRS Web Sémantique, 2002.

[Denoue, 00] L. Denoue : *De la création à la capitalisation des annotations dans un espace personnel d'informations*. Thèse de doctorat en Informatique, Université de Savoie, soutenue le 26 octobre 2000.

[Desmoulins, 02] Desmoulins.C, Grandbastien.M: *Des ontologies pour la conception de manuels de formation à partir de documents techniques*. Science et techniques éducatives, Hermès, Paris.2002.

[Firan, 07] S. Firan, W. Nejdl, R. Paiu: *The Benefit of Using Tag-Based Profiles.* Proceedings of the 2007 Latin American Web Conference LA-WEB, page 32-41. Washington, DC, USA, IEEE Computer Society, (2007)

[Gerald, 02] Gerald. J, Kowalski.M, T. Maybury: *Information storage and retrieval systems Theory and Implementation*, Second Edition, KLUWER ACADEMIC PUBLISHERS, 2002.

[Gruber, 07] Gruber.T : *Ontology of folksonomy: A mash-up of apples and oranges*. Int. Journal on Semantic Web and Information Systems, 2007.

[Guy, 06] Guy.M, Tonkin.E : *Folk-sonomies. Tidying up Tags?* D-lib Magazine Volume 12, N° 1, 2006.

[Golder, 05] Golder Scott A and Bernardo A. Huberman: *The Structure of Collaborative Tagging Systems*. Journal of Information Science32(2):198--208Aug, 2005.

[Halpin, 07] Halpin H., Robu.V, Shepherd.H. *The Complex Dynamics of Collaborative Tagging*. In WWW : ACM Press, 2007.

[Handschuh, 03] Handschuh.S, Staab.S : *Annotating of the shallow and the deep web*. Annotation for the semantic web. Amesterdam, IOS press, 2003.

[Heck ,99] Heck R. M., Luebke S. M., Obermark C. H. (1999). *A Survey of Web Annotation Systems*. Rapport interne, Dep. Of Mathematics and Computer Sience, Grinnell College, USA.

[Heckner, 08] Heckner.M, Neubauer.T, and Wolff.C: *Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types*. SSM '08: Proceeding of ACM workshop on Search in social media, New York, USA, 2008.

[Huang, 08] Huang.Y, C. Hung, J.Hsu : *You are what you tag* : Association for the Advancement of Artificial Intelligence (www.aaai.org), 2008.

[Huart, 96] Huart, P : *Définition d'un poste de lecture active de documents électroniques*, Rapport de stage, ENIB-ENSEEIHT-IRIT, Toulouse, 1996.

[Jaillet,03] Jaillet.S, Teisseire.M, Dray.G. *Adéquation des modèles de représentation aux méthodes de catégorisation*. LIRMM-CNRS – ISIM - Université Montpellier, 2003.

[Kahan, 02] Kahan.J, Koivunen.M, Prud'Hommeaux.E, Swick.R: *Annotea: An Open RDF Infrastructure for Shared Web Annotations*, in Proceedings of the WWW10 International Conference, Hong Kong, Mai 2001.

[Kalyanpur, 04] Kalyanpur. A , J.Golbeck , J.Hendler , B.Parsia: *SMORE - Semantic Markup, Ontology, and RDF editor*. In Proceedings of the 3rd International web semantic conference (ISWC), Japan (poster), 2004.

[Koivunen, 03] Koivunen.M., Swick.R, Kahan.J, Prud'hommeaux.E. *Annotea shared Bookmarks*. Proceedings of the KCAP 2003 workshop on knowledge markup and semantic annotation, Sanibel, Florida-USA, octobre 2003.

[Kichou, 11] S.Kichou, H.Mellah, Y.Amghar: *Weighting Tags Approach Based on User Profil.* International Conference on Active Media Technology (AMT 2011), expected on September 7-9, 2011, Lanzhou, China.

[Lechani, 05] L. Lechani-Tamine, M. Boughanem *: Accès personnalisé à l'information : Approches et Techniques,* IRIT : Institut de Recherche en Informatique de Toulouse, Equipe SIG/RFI, Rapport interne, Janvier 2005.

[Le Deuff, 06] Le Deuff.O*: Folksonomies, Les usagers indexent le web.* BBF 2006 - Paris, t. 51, n° 4.

[Liang, 09] Liang, Huizhi and Xu, Yue and Li, Yuefeng and Nayak, Richi (2009) *Tag based collaborative filtering for recommender systems*. In: Proceedings of Rough Sets and Knowledge Technology: 4th International Conference, 14-16 July 2009.

[Limpens, 08] F.Limpens, F.Gandon1, M.Buffa: *Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées : un état de l'art*. 19es Journées francophones d'ingénierie des connaissances, Nancy, 2008.

[Limpens, 09] F.Limpens, F.Gandon1, M.Buffa *: Sémantique des folksonomies: structuration collaborative et assistée.* Ingénierie des Connaissances, Hammamet : Tunisie, 2009.

[Marshall, 97] Marshall.C: *Annotation: from paper books to digital library*. Proceedings of second ACM international conference on digital libraries. ACM press Newyork, USA, 1997.

[Marshall, 98] Marshall.C: *Toward an ecology of hypertext annotation*. Proceedings of the ninth ACM conference on hypertext and hypermedia. ACM press Newyork, USA, 1998.

[Marlow, 06] Marlow.C, Mor N, Danah B, and Marc.D. *Tagging, taxonomy, flickr, article, toread*. In Collaborative Web Tagging Workshop at WWW'06, Edinburgh, UK, 2006.

[Mathes, 04] Mathes.A: Folksonomies - *Cooperative Classification and*

Communication Through hared Metadata. Rapport interne, GSLIS, Univ. Illinois Urbana- Champaign, 2004.

[Medin, 89] D. L. Medin. *Concepts and conceptual structure*. American Psychologist, 44(12):1469-1481, 1989.

[Mika, 05] Mika.P: *Ontologies are Us: a Unified Model of Social Networks and Semantics*. In ISWC, volume 3729 of LNCS, p. 522–536: Springer. 2005.

[Mille, 05] Mille.D: *Modèles et outils logiciels pour l'annotation sémantique des documents pédagogiques.* Thèse de doctorat en Informatique, Université Joseph Fourier – Grenoble, 2005.

[Nie,01] Nie J.Y, *Le domaine de recherche d'information – Un survol d'une longue histoire*, Département d'informatique et recherche opérationnelle Université de Montréal, 2001

[Paradis, 96] Paradis.F : *Un modèle d'indexation pour les documents textuels structurés*. Thèse de doctorat de l'Université Joseph Fourier - Grenoble 1, 1996.

[Passant, 07] Passant. A (2007). *Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs*. In Int. Conf. Weblogs and Social Media.

[Passant, 08] A.Passant, P.Laublet: *Ontologies pour le Web 2.0*: 19es Journées francophones d'ingénierie des connaissances, Nancy, 2008.

[Peccatte, 02] P.Peccatte : *Les métadonnées, un élément clé de la gestion de contenu*. ATICA-Deuxième journée de la réutilisation des données, octobre 2002.

[Prié, 04] Y.Prié, S.Garlattit : *Métadonnées et annotation dans le web sémantique*. Hors série 2004-web sémantique. Revue en science du traitement de l'information, 13 (information-Interaction-Intelligence), cépaduès, Toulouse, France, 2004.

[Reeve, 05] L.H.Reeve, H.Han: *Survey of semantic annotation* .SAC 2005, ACM Press, ISBN 1-58113-964-0, Santa-fe NY USA, mars 2005.

[Roussey, 01] Roussey C. (2001). *Une méthode d'indexation sémantique adaptée aux corpus multilingues.* Thèse de doctorat, Institut National des Sciences Appliquées (INSA) de Lyon, Dec. 2001.

[Rupert, 10] Rupert.M, S. Hassas : *Building Users' Profiles from Clustering Resources in Collaborative Tagging Systems*. AMT'10, Proceedings of the 6th international conference on Active media technology, 2010.

[Salton, 70] Salton G. (1970). *The SMART retrieval system: Experiments in automatic document processing*. Prentice Hall.

[Sauvagnat, 05] Sauvagnat K. *Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés*, Thèse de Doctorat de l'Université Paul Sabatier, 2005.

[Specia, 07] Specia.L, Motta.E: *Integrating folksonomies with the semantic web*. 4th European Semantic Web Conference, 2007.

[Tamine, 06] Tamine-Lechani.L, N.Zemirli, W.Bahsoun: *Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'informations*. Actes de la Conférence francophone en Recherche d'Information et Applications (CORIA 2006), Lyon : France 2006.

[Tebri,04] Tebri H. *Formalisation et spécification d'un système de filtrage incrémental d'information*. Thèse de doctorat de l'université Paul Sabatier, Toulouse, 2004.

[Vanderwal, 05] Vanderwal.T: *Explaining and Showing Broad and Narrow Folksonomies.* http://www.vanderwal.net/random/entrysel.php?blog=1635, *2005.*

[Veron, 97] M. Veron : *Modélisation de la composante annotative dans les documents électroniques.* Master de recherche, IRIT,Toulouse. P.53, 1997.

[Wang, 09] Wang.J, M. Clements , J. Yang , A .de Vries, Marcel J.T. Reinders *Personalization of tagging systems*. Information Processing & Management, 2009.

[Wasserman, 94] Wasserman.S., & Faust, K. (1994). *Social Network Analysis*. Cambridge, UK: Cambridge University Press, 1994.

[Xu, 08] Xu.S, S.Bao, B. Fei: *Exploring Folksonomy for Personalized Search*. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008.

[Xu, 06] Xu.Z., Y. Fu, J. Mao, and D. Su. *Towards the semantic web:*

Collaborative tag suggestions. WWW Workshop on Collaborative Web Tagging, 2006.

Annexe A : Exemples d'outils d'annotation

1- Introduction

Cette annexe est consacrée à la description d'un certain nombre d'outils d'annotation que nous avons évoqué dans le premier chapitre. Nous détaillons dans la section 2 quelques exemples de ces outils. La section 3 est consacrée aux langages utilisés pour les annotations sémantiques.

2- Les outils d'annotation

- **Annotea et Amaya**

Annotea **[Koivunen, 03]**, un projet LEAD[6] (Live Early Adoption and Demonstration) du W3C[7], pour l'annotation de documents Web. Il met en évidence la collaboration par le partage des annotations. Annotea utilise RDF pour la description des annotations, XPointer pour leurs localisations. Les annotations peuvent être enregistrées soit localement ou bien dans un ou plusieurs serveurs RDF public. Amaya est la première implémentation d'Annotea.

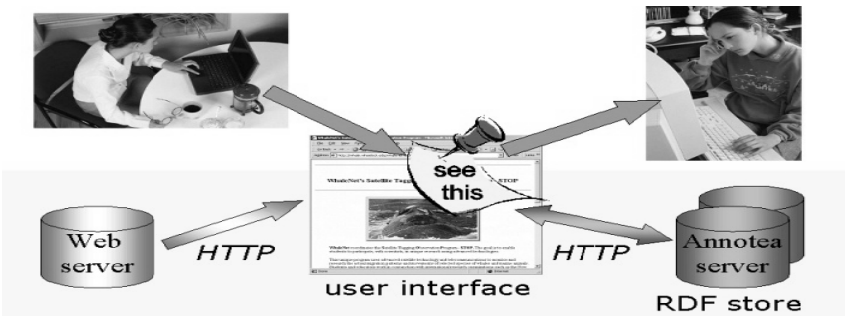

Fig 01 : *Architecture d'Annotea* ***[Koivunen, 03]***

[6] http://www.w3.org/2001/Annotea/

[7] : Le World Wide Web Consortium est un organisme de standardisation à but non-lucratif, fondé en octobre 1994 comme un consortium chargé de promouvoir la compatibilité des technologies du World Wide Web telles que HTML,XHTML, XML, RDF, CSS, PNG, SVG et SOAP. Le W3C n'émet pas des normes au sens européen, mais des recommandations à valeur de standards industriels.

Amaya est un Navigateur/éditeur de page web, développé conjointement par l'INRIA[8] et le W3C, dont les deux principaux objectifs sont : la démonstration des nouvelles technologies web.

Amaya suit plusieurs recommandations W3C (XML, XHTML, MathML, SVG…) et la création de page web conforme à ces recommandations. Il permet le partage de métadonnées, de marques pages et d'annotations textuelles sur des pages Web.

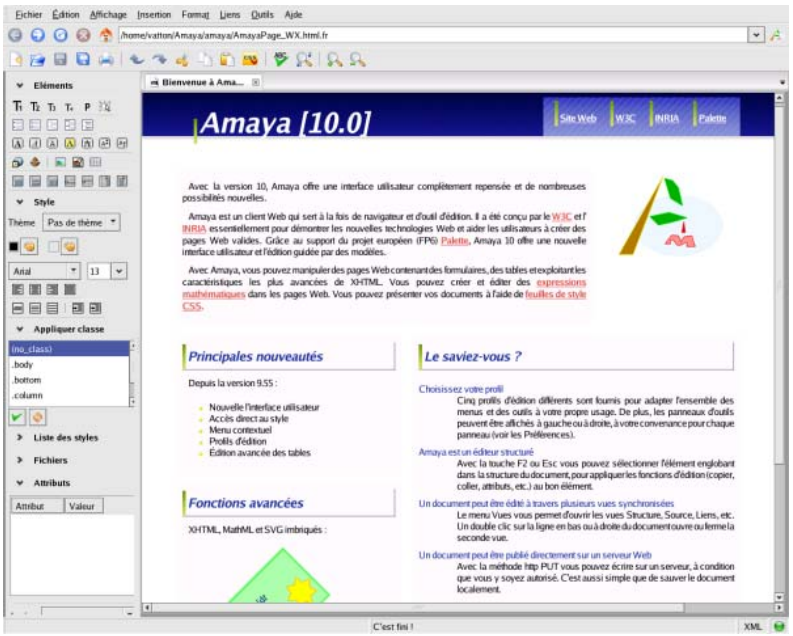

Fig02: *Site officiel d'Amaya*

Interface d'ajout d'une annotation:

- Choix de la cible en surlignant ;
- Choix de l'option « annotate selection » dans le menu « annotation » ;
- Ouverture d'une fenêtre pop-up pour saisir le contenu de la note.

Interface de visualisation d'une annotation :

- Icône stylo sur lequel on clique pour faire apparaître une fenêtre contenant le contenu de l'annotation ;
- S'il s'agit d'une réponse à une annotation (reply sur une annotation), visualisation dans une file de discussion (liste des reply).

- **Annozilla**

Annozilla permet l'annotation de page web, développé pour le navigateur Mozilla-Firefox. Il se présente sous la forme d'un plug-in Open Source (Javascript, XUL, XPCOM,

[8] : Institut National de Recherche en Informatique et en Automatique, fondé en 1967.

HTML) respectant le standard Annotea. Il permet de voir et de créer des annotations associées à une page Web. IL utilise RDF et la technologie XPointer (basé sur XML) pour permettre de localiser les annotations sur la page Web.

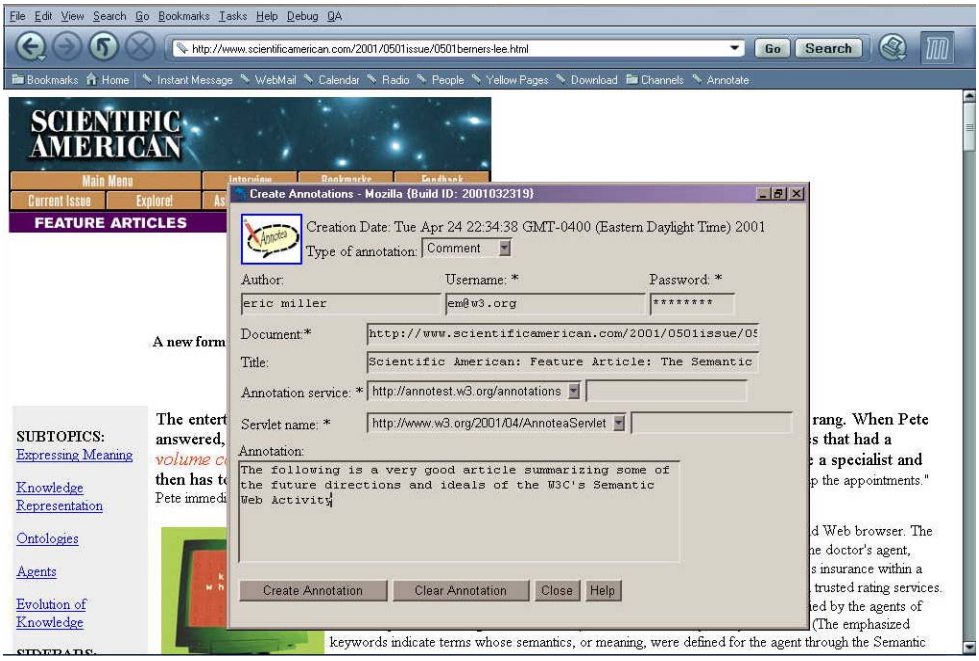

Fig 03: *L'annotation avec Annozilla*

- **IMarkup**

IMarkup est un plug-in d'annotation de pages web, pour Internet Explorer.

Pour ajouter une annotation il faut choisir entre différents types d'annotation proposés par l'outil : post-it, marque typographique, schéma réalisé avec un pinceau. Pour ajouter un commentaire au clavier il suffit de cliquer sur l'objet inséré.

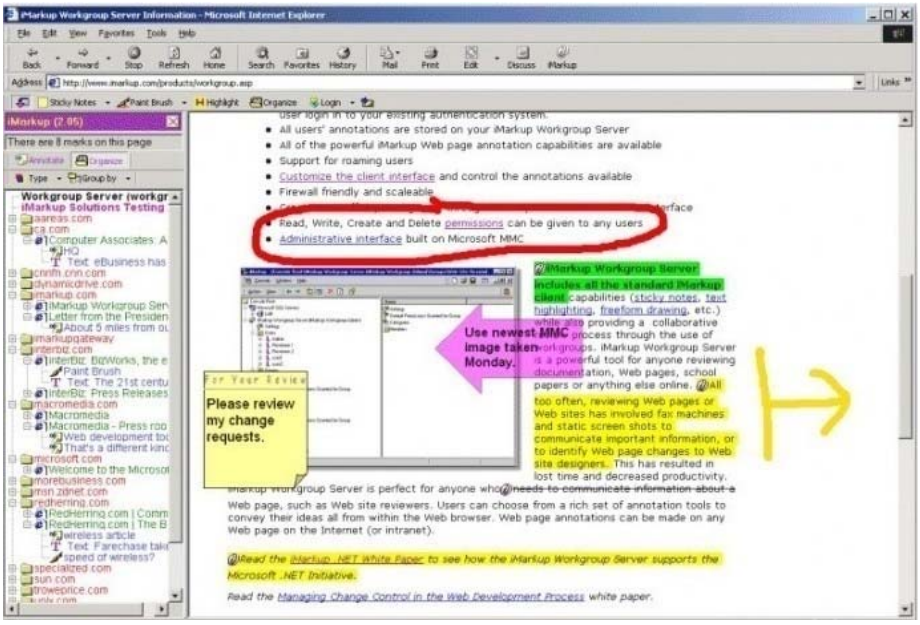

Fig 04: *L'annotation avec Imarkup*

- **XLibris**

XLibris est un outil dédié à l'annotation de documents numériques de manière intuitive via des outils « à main levée ». Il aide les utilisateurs à lire les documents numériques de manière très semblable à la lecture de documents papiers, en donnant la possibilité d'annoter manuellement sur le document via un stylet : stylo coloré, surligneur, gomme et différentes commandes pour marquer, dessiner et écrire sur une page.

Fig 05: *Annoter avec XLibris*

- **Connotea**

S'adressant à la communauté scientifique, Connotea est un outil d'organisation et de partage de références bibliographiques. Il permet l'annotation de papiers scientifiques et leurs organisations à travers un profil personnel. L'utilisateur peut ajouter un document à sa bibliothèque ou à une bibliothèque partagée, et peut ajouter ses propres annotations (tags).

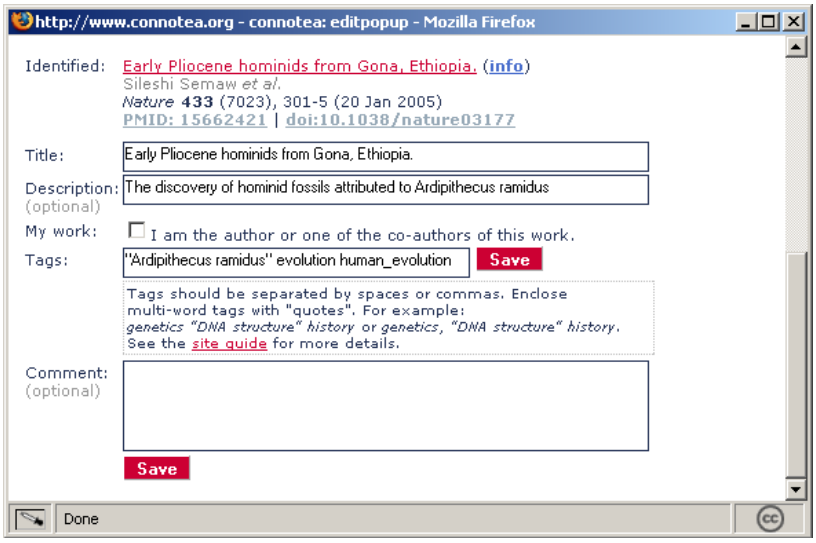

Fig 06 : *Fenêtre d'annotation dans Connotea*

Ces quelques outils cités ci-dessus, à leurs différences, permettent tous la création, la visualisation et le partage des annotations. La création se fait soit en surlignant (Amaya), ou en déposant l'objet annotation sur la partie à annoter (IMarkup). La visualisation des annotations peut se faire sur une info-bulle en passant sur la cible, sur une fenêtre pop-up lorsqu'on clique sur l'objet annoté, ou une visualisation directe sur le document (sur les marges par exemple).

Nous avons présenté dans le chapitre I, section 6 une classification des outils d'annotation faite par **[Azouaou, 05]** en se focalisant sur les types de l'objet et l'activité annotation. Entre autres, **[Heck ,99]** présente un comparatif de quelques systèmes d'annotation, basé sur la localisation de l'annotation, de sa destination à un groupe privé ou public et sur la possibilité ou non de la consulter et de la rechercher. **[Prié, 04]** se base sur le type de ressources annotées, le langage d'annotation, l'architecture et l'utilisation des annotations. Quant à **[Reeve, 05]**, il présente une synthèse et une classification, des plateformes d'annotations semi-automatiques (Armalido, kim, MnM, OntoMAT, SenTag)

3- Outils et langages d'annotations sémantiques

[Benna, 08] a fait une synthèse sur les outils d'annotation sémantique les plus référencés, en plus d'Annotea dont nous avons parlé, l'auteur a cité les outils suivants : COHSE[9] Conceptual Open Hypermedia Services Environment, KIM[10], MnM[11], Ontomat Annotizer[12] , SHOE Knowledge Annotator[13], YAWAS et METEOR-S.

L'auteur a également fait une étude qui a permis de conclure que les systèmes d'annotation sémantique varient dans leur architecture, les outils d'extraction de l'information et dans les méthodes et les langages d'annotations. Ils dépendent aussi du cadre de leur utilisation (destiné à une collaboration, une recherche, une intégration,…). Ils peuvent être destinés à être traités par l'humain, l'agent logiciel ou par les deux.

L'expression des annotations sémantiques se fait principalement par deux langages, RDF et Topic Maps, mais il existe d'autres langages tels que HTML-A, SHOE qui sont des extensions de HTML permettant l'insertion des annotations sémantiques. OWL est utilisé pour décrire les ontologies dont les concepts sont utilisés pour l'annotation et RDF-S pour décrire des ontologies légères (utilisées dans le Tagging).

[9] http://cohse.semanticWeb.org
[10] http://www.ontotext.com/kim/
[11] http://kmi.open.ac.uk/projects/akt/MnM/
[12] http://annotation.semanticWeb.org/ontomat/index.html
[13] http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html

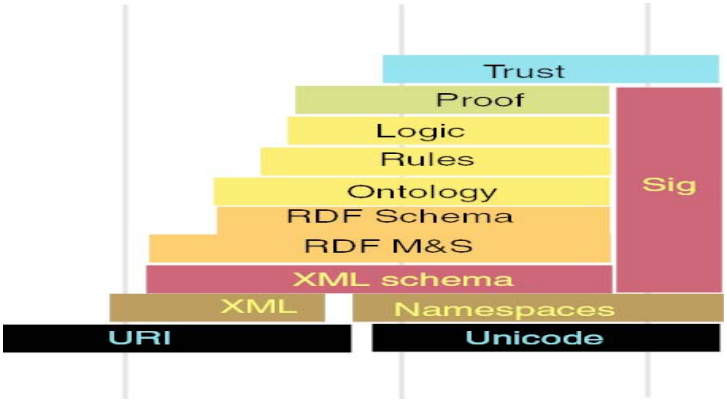

Fig 07 : *Langages du Web Sémantique*

Nous définissons RDF, brièvement Topic maps, RDF-S puis OWL.

a- ***Resource Description Framework (RDF) :*** *est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de manière à permettre le traitement automatique de telles descriptions. Développé par le W3C, RDF est le langage de base du Web sémantique,* ***[Attar, 04].***

- ***Objectif***

L'objectif de RDF est de proposer un cadre formel de définition de Métadonnées, en ce sens, RDF est un langage formel spécialisé dans les Métadonnées. Son objectif est de rendre plus pertinent le traitement automatisé des informations contenues sur le Web, par la possibilité de fournir aux outils de traitement une information plus sémantique que les seuls mots contenus dans un document. RDF permet d'annoter toute ressource du web qui possède une adresse URI.

- **Principe**

La structure fondamentale de toute expression en RDF est une collection de triplets, chacun composé d'un sujet, un prédicat et un objet. Un ensemble de tels triplets est appelé un graphe RDF. Ceci peut être illustré par un diagramme composé de nœuds et d'arcs dirigés, dans lequel chaque triplet est représenté par un lien nœud-arc-nœud. Le triplet est appelé assertion, **[Peccatte, 02]**.

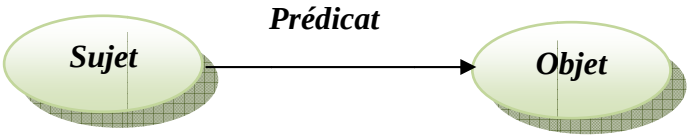

Fig 08 : *Structure d'une assertion*

- Le sujet représente la ressource à décrire, identifié par une URI ;
- Le prédicat représente un type de propriété applicable à cette ressource, ce prédicat est aussi identifié par une URI ;
- L'objet représente une donnée ou une autre ressource : c'est la valeur de la propriété.

Exemple

Le document **RadioTV-NewsML** in Japan **a pour auteur M.Onishi**

Sujet **Prédicat** **Objet**

Syntaxe d'un document RDF

Un document RDF peut être représenté sous forme de graphe, comme illustré dans la figure 05, cependant cette représentation n'est pas interprétable par la machine, d'où représentation (sérialisation) dans un document XML (langage RDF/XML), (voir figure06). Celui-ci présente l'avantage d'être standardisé et interopérable.

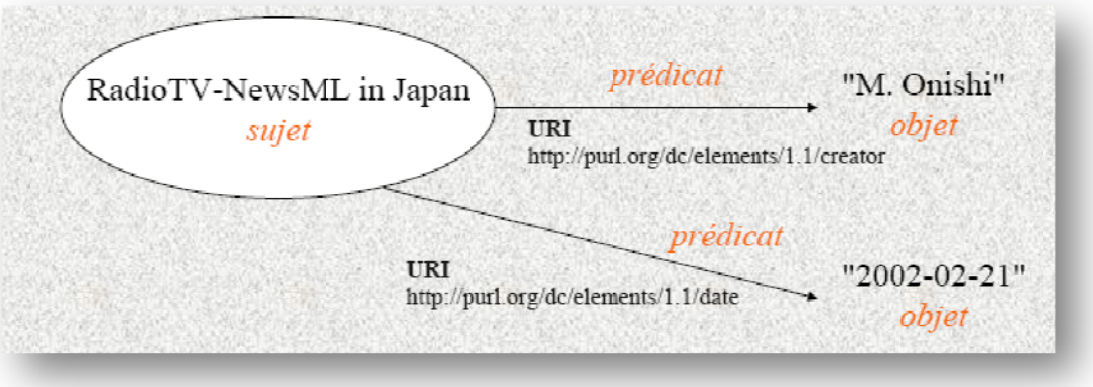

Fig 09: *Graphe RDF*

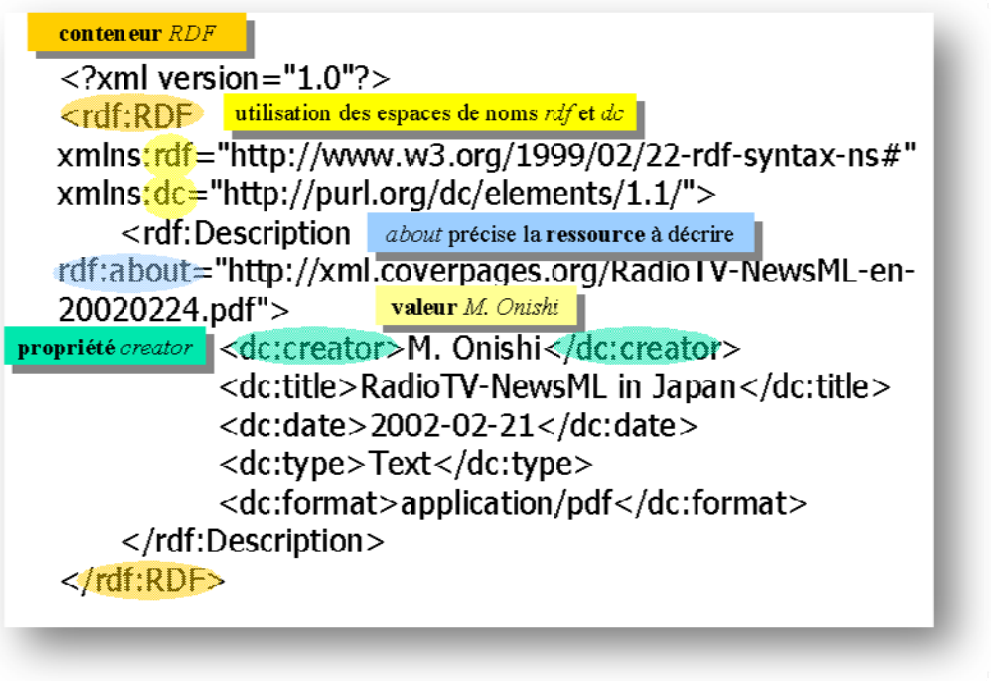

Fig 10 : *Version XML du graphe RDF*

b- RDF-S (RDF Schéma) : est une extension de RDF, il précise la notion de propriété définie par RDF en permettant de donner un type ou une classe au sujet et à l'objet des triplets. Il fournit un mécanisme permettant de spécifier les classes dont les ressources seront des instances, **[Boufaida, 08].**

- **Les classes de RDF-S :** les classes de base définies comme une partie du vocabulaire de RDF-S :
  - rdfs : Ressource : toutes les ressources décrites dans des expressions RDF sont des instances de cette classe ;
  - rdfs : Property : est le sous ensemble des ressources qui sont des propriétés (prédicats) ;
  - rdfs : Class : permet de déclarer une ressource RDF comme classe pour d'autres ressources.
- **Les propriétés de RDF-S :** les propriétés suivantes sont des instances de la classe rdfs : Property, elles expriment les relations entre les classes et leurs instances
  - rdfs : Type : permet d'indiquer qu'une ressource appartient à une classe et possède toute les caractéristiques qu'une autre ressource appartenant à cette classe doit avoir ;
  - rdfs : SubClassOf : spécifie des hiérarchies de classes (sous classes) ;

- rdfs : SubPropertyOf : indique qu'une propriété est une spécialisation d'une autre.

Les contraintes de RDF-S

- rdfs :ConstraintProperty :
- rdfs : Range : définit la classe ou le type de données des valeurs de la propriété (où la relation arrive).
- rdfs :Domain : définit la classe des sujets liée à une propriété (d'où la relation part).

La sémantique sur le Web est très riche. C'est évident que les primitives offertes par RDFS sont insuffisantes pour la modélisation dans le Web Sémantique. D'où l'apparition d'OWL.

c- Les cartes topiques (Topics Maps) : sont un outil très général de représentation des connaissances, dont le but est d'agréger autour d'un point unique d'indexation (appelé topic) toutes les informations disponibles concernant un sujet donné, et de relier ces points par un réseau sémantique de relations appelées associations.

Concept : Un topic map représente une information en utilisant des « sujets » (topics) qui représentent tout concept, tel qu'une personne, un groupe de personnes, une couleur, un pays, une organisation, un module logiciel, un fichier individuel, des évènements, en utilisant des « associations » qui représentent les relations entre ces « sujets », et des « occurrences » qui représentent des relations entre des sujets et des ressources informationnelles qui s'y rapportent, [wikipédia][14]**.**

d- OWL (web ontology language) : est un langage ontologique du web qui est une extension de RDF-S pour définir et instancier les ontologies du web. C'est un dialecte XML basé sur une syntaxe RDF, il fournit trois sous langages OWL Lite, OWL DL, OWL Full, **[Boufaida,08]**. OWL est une logique de description conçue en dessus de RDF, il permet de décrire des ontologies en définissant des terminologies qui permettent de décrire des domaines de connaissances. **[Azouaou, 07]**.

[14] http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

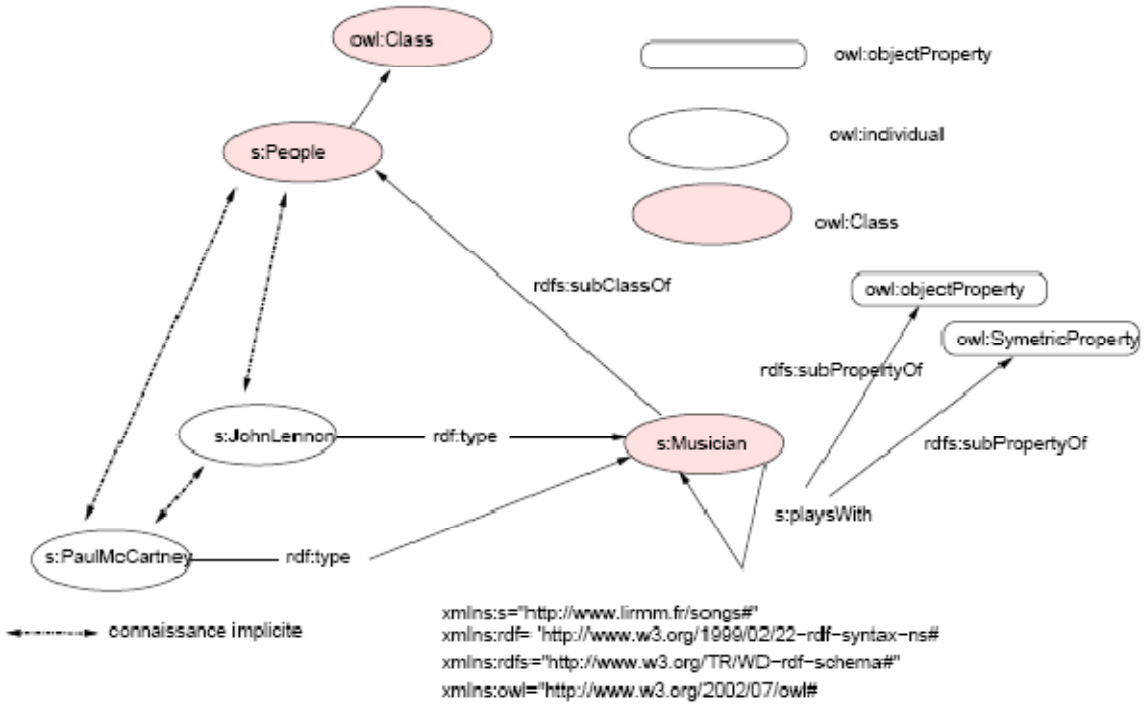

Fig 11 : *Illustration OWL*

Annexe B : Exemples de Systèmes du Tagging Collaboratif

Il existe sur le web une multitude de sites permettant le Tagging collaboratif. Dans les travaux **[Marlow, 06],** Cameron Marlow a choisi de citer douze exemples qu'elle estime représentatifs de la diversité du point du vue types de ressources à tagguer, architecture et motivations d'utilisateurs.

- ***Del.icio.us (http://del.icio.us):*** site de 'social bookmarking', dit signet social ou marque-pages social, permettant le Tagging de pages web et sites. Une liste de tags populaires est affichée. (Fig 12)

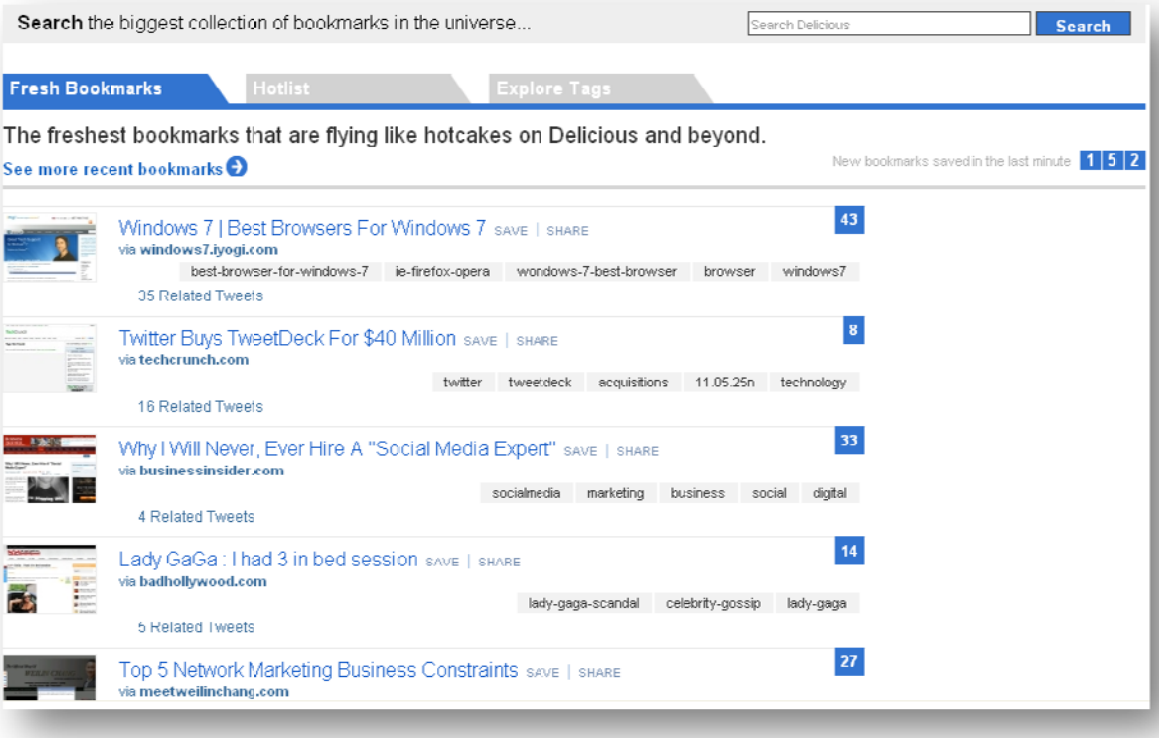

Fig 12 : *Page d'accueil de Delicious*

- ***Yahoo! MyWeb2.0 (http://myweb.yahoo.com):*** similaire à Del.icio.us, tout en incluant un réseau social des contacts.
- ***CiteULike (http://www.citeulike.org/):*** site permettant le Tagging des citations et des références : papiers scientifiques ou livres.

- ***Flickr*** (***http://www.flickr.com***): system de partage de photos, permettant à un utilisateur de sauvegarder et tagguer ses photos et aussi celles des autres utilisateurs (Fig 13).

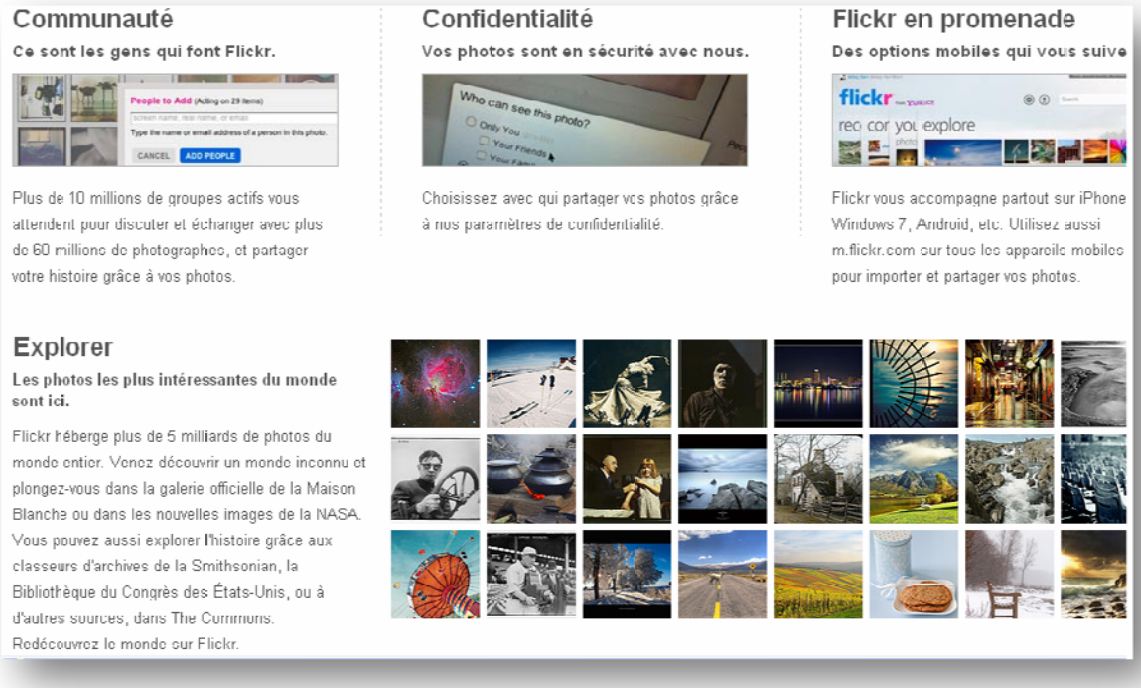

Fig 13 : *Page d'accueil de Flickr*

- ***YouTube (http://www.youtube.com):*** system de partage de vidéos, il permet de consulter et décrire les vidéos.
- ***ESP Game (http://www.espgame.org/) :*** un jeu de Tagging sur internet où les utilisateurs sont couplés aléatoirement, et essayent de deviner les tags qu'utilise l'autre pour décrire la même ressource (généralement une photo).
- ***Last.fm (http://www.last.fm):*** une base de données de musique, permet de tagguer les artistes, albums et chansons.
- ***Yahoo! Podcasts (http://podcasts.yahoo.com/):*** un site qui indexe des *podcasts* (diffusion de contenus audio), et permet aux utilisateurs de les tagguer.
- ***Odeo (http://www.odeo.com/):*** un autre site de *podcast* permettant le Tagging et la recherche.
- ***Technorati (http://www.technorati.com/):*** est un agrégateur de blogs, et un outil de recherche de ceux-ci. Permettant leurs Tagging par leurs auteurs.

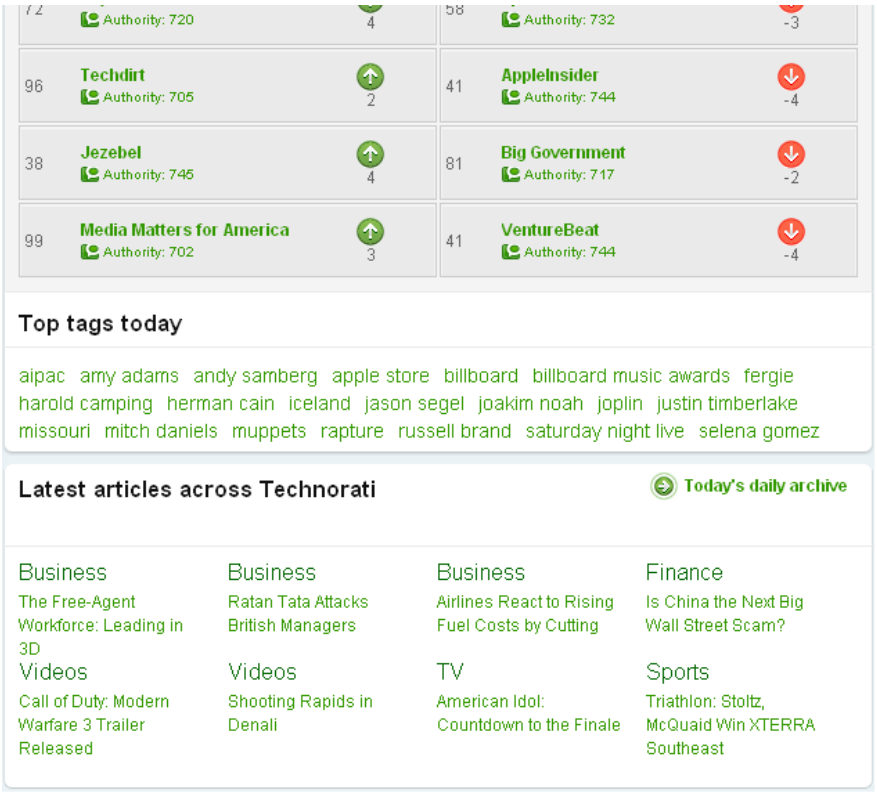

Fig 14 : *La rubrique tags et blogs populaires dans Technorati*

- ***LiveJournal (http://www.livejournal.com/):*** un blog et site communautaire permettant aux utilisateurs d'éditer leurs profils personnels, et de créer leurs blogs.
- ***Upcoming (http://upcoming.org/):*** site permettant aux utilisateurs d'introduire des événements futurs (concerts, pièces de théâtre, expositions…) et les tagguer.

Annexe C : L'ontologie WordNet

1- Introduction

Nous avons utilisé dans notre travail l'ontologie WordNet pour la récupération des profondeurs de tags. Ces profondeurs rentrent dans le calcul de l'expertise de l'utilisateur. Nous consacrons cette annexe pour la description de WordNet, et la définition de quelques notions liées à cette ontologie.

2- Description

WordNet est une base de données lexicale développée, sous la direction de *George A. Miller*, par des linguistes du laboratoire des sciences cognitives de l'université de Princeton (site officiel WordNet)[15]. Elle a été initialement conçue dans le cadre d'un projet lancé en 1985 et gracieusement financé par l'agence de renseignements américaine (CIA), avec l'objectif de tester les déficits lexicaux dans des expériences de psychologie cognitive. A l'origine, ces concepteurs ne prétendaient construire ni une structure conceptuelle, ni une ontologie, mais bien une ressource lexicale rendant compte de l'usage des mots et de leur mise en relation dans la langue, **[Baziz, 05]**.

Le système se présente sous la forme d'une base de données électronique qu'on peut télécharger (WordNet est distribuée avec une licence spéciale très libérale, permettant de l'utiliser commercialement ou à des fins de recherche) sur un système local et y accéder à partir d'un programme à l'aide d'interfaces disponibles pour de nombreux langages de programmation.

Concernant le contenu de WordNet, il couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise. WordNet a un réseau de 144 684 termes organisés en 109 377 noeuds (concepts) appelés Synsets.

[15] http://wordnet.princeton.edu/

Catégories	Mots	Concepts	Total Paires Mot-Sens
Nom	107 930	74 488	132407
Verbe	10 806	12 754	23255
Adjectif	21 365	18 523	31077
Adverbe	4 583	3 612	5721
Total	144 684	109 377	192460

Tableau 01: *Le nombre de mots et de concepts dans WordNet,* ***[Baziz, 05]***

2.1 Le Concept

Selon le dictionnaire de l'académie française, *"Le concept regroupe les objets qu'il définit en une même catégorie appelée « classe » "*, **[Baziz, 05]**. De manière générale, le terme concept est souvent utilisé comme se référant à toute notion, de l'idée au lexème, en passant par l'entité et la catégorie. Selon **[Medin, 89]**, globalement un concept est une idée qui inclut tout ce qui est caractéristiquement associé à elle.

2.2 Les Synsets

La composante atomique sur laquelle repose le système entier est le synset (synonym set), un groupe de mots interchangeables, dénotant un sens ou un usage particulier.

D'après **[Baziz, 05], l**e Synset contient trois parties:

• Le terme représentant du Synset : c'est le terme pour lequel le concept (synset) est identifié. Dans WordNet, les termes les plus utilisés sont placés en premier,

• Les termes synonymes : une liste de termes interchangeables séparés par des virgules,

• Le glossaire : il est mis entre parenthèses et vient après le symbole "--". Il contient une définition du concept avec éventuellement un ou plusieurs exemples du monde réel (mis entre double côtes, "").

Exemple: Le mot ' java' a 3 sens dans WordNet:

1. (2) **Java** -- (an island in Indonesia south of Borneo; one of the world's most densely populated regions)
2. (1) coffee, **java** -- (a beverage consisting of an infusion of ground coffee beans; "he ordered a cup of coffee")
3. **Java** -- (a simple platform-independent object-oriented programming language used for writing applets that are downloaded from the World Wide Web by a client and run on the client's machine)

2.3 Les relations sémantiques dans WordNet

Les relations sémantiques existantes dans WordNet peuvent être résumées comme suit :

- Relation ***Synonymie***, les synonymes étant associés à la classe Concept.

- Relation ***Hyperonymie:*** C'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques. Y est un *hyperonyme* de X si X est un type de (kind of) Y.
- Relation ***Hyponymie***: C'est le terme spécifique utilisé pour désigner un membre d'une classe (relation inverse de Hyperonymie). X est un hyponyme de Y si X est un type de (kind of) Y.
- Relation ***Holonymie***: Le nom de la classe globale dont les noms méronymes font partie. Y est un holonyme de X si X est une partie de (is a part of) Y.
- Relation ***Méronymie***: Le nom d'une partie constituante (part of), substance de (substance of) ou membre (member of) d'une autre classe (relation inverse de l'holonymie). X est un méronyme de Y si X est une partie de Y. exemple : {voiture} a pour méronymes {{porte}, {moteur}}.

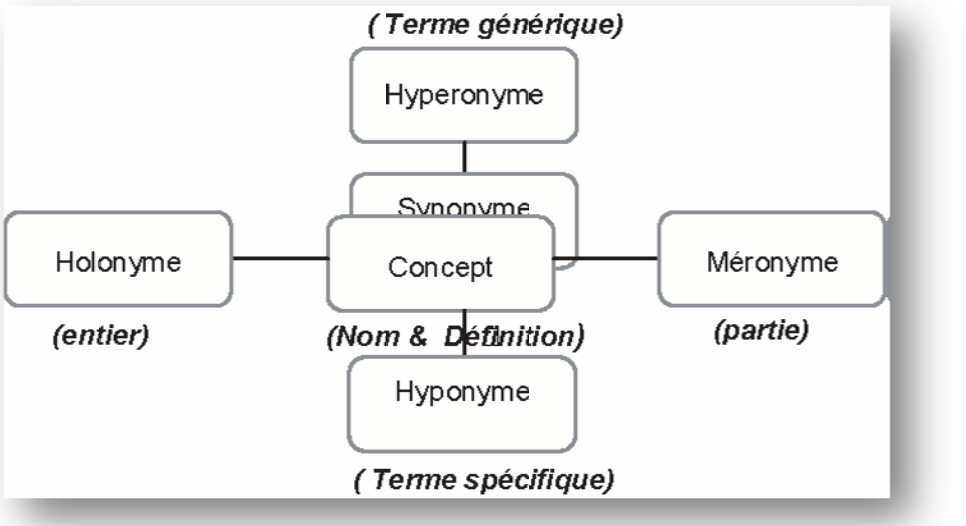

Fig 15 : *Les principales relations sémantiques dans WordNet,**[Baziz,05]***

Annexe D : Indexation et Recherche d'information

1. Introduction

La Recherche d'Information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information **[Tebri,04].** Elle remonte au début des années 1950 et le nom de «recherche d'information» (*information retrieval*) fut donné par Calvin N. Mooers en 1948 pour la première fois quand il travaillait sur son mémoire de maîtrise. La première conférence dédiée à ce thème – *International Conference on Scientific Information* - s'est tenue en 1958 à Washington **[Nie, 01]**.

La RI a commencé à se développer dans les années 60, principalement à travers la problématique de recherche documentaire où on s'est intéressé à l'accès à l'information dans des bibliothèques. A la fin des années 60 et au début des années 70, G. Salton a développé le système SMART *(Salton's Magical Automatic Retriever of Text)* **[Salton, 70]** qui a beaucoup influencé le domaine.

2. Processus de recherche d'information

2.1 Architecture du processus

Le processus de Recherche d'Information a pour but la mise en relation des informations disponibles dans le fond documentaire d'une part, et les besoins en information des utilisateurs d'autre part. Ces besoins sont formalisés par l'utilisateur sous forme de requêtes. La mise en relation des besoins utilisateurs et des informations est effectuée grâce à un Système de Recherche d'Information (SRI) dont le but est de retourner à l'utilisateur le maximum de documents pertinents par rapport à son besoin (et le minimum de documents non pertinents). Pour cela, le système de recherche d'information met en relation les informations disponibles (les documents du corpus) d'une part et les besoins de l'utilisateur (la requête utilisateur) d'autre part.

Le processus de recherche, appelé souvent processus en U (fig16) assure essentiellement des opérations de représentation et d'interrogation. La représentation (ou indexation) permet d'extraire à partir des documents et des requêtes des représentations qui couvrent au mieux leurs contenus sémantiques. L'interrogation (ou la recherche) représente le noyau du système

et comprend les fonctions de décision permettant d'associer à chaque requête l'ensemble des documents pertinents à restituer.

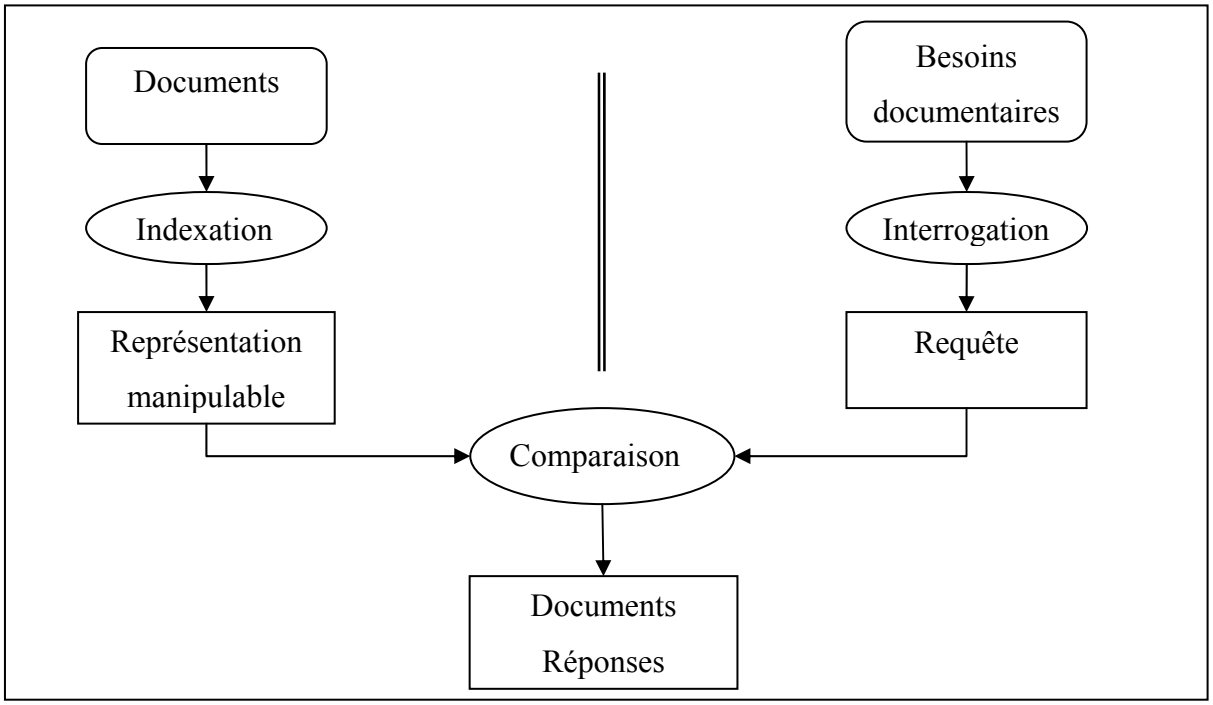

Fig 16 : *Processus de recherche d'information*

2.2 Indexation

[Roussey, 01] en parle, "*L'indexation consiste à identifier l'information contenue dans tout texte et à la représenter au moyen d'un ensemble d'entités appelé index pour faciliter la comparaison entre la représentation d'un document et d'une requête*."

Dans les systèmes classiques de recherche d'informations, l'indexation est organisée en trois étapes : l'Extraction, la sélection et la pondération des termes. (Fig 17).

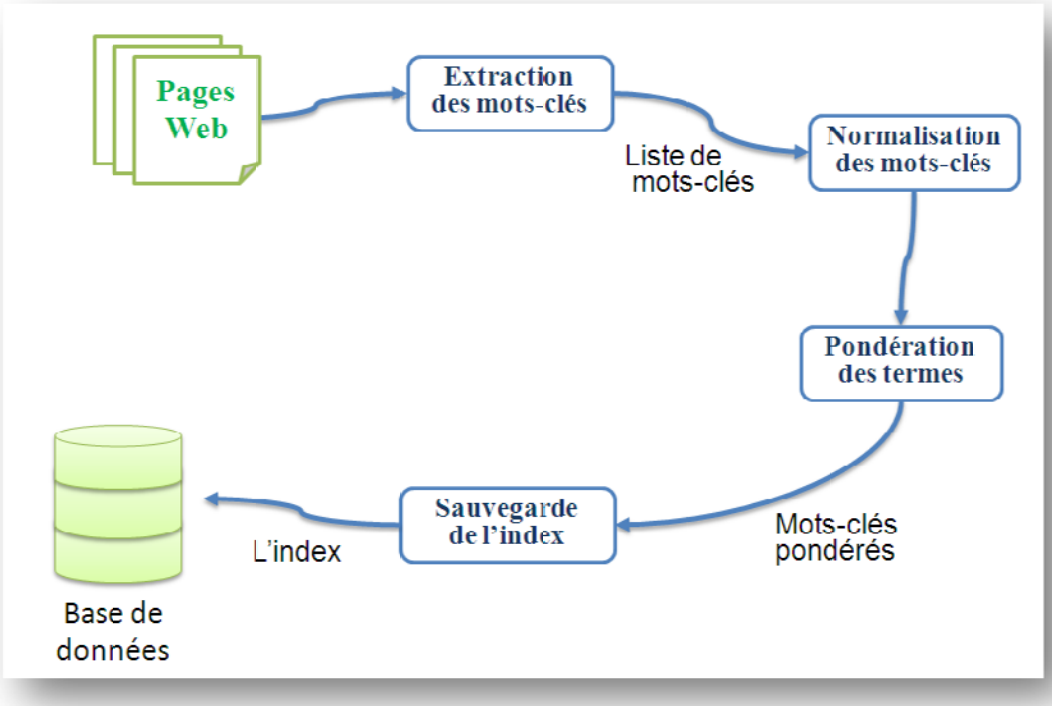

Fig 17 : *Processus d'indexation,* ***[Dahak, 06]***

Le processus d'indexation est mis en œuvre afin d'extraire préalablement une représentation homogène du contenu sémantique, sous forme de termes d'indexation qui sont des éléments d'un langage d'indexation.

Les étapes du processus d'indexation sont présentées ci-dessous.

2.1.2 L'extraction des mots clés

Cela permet d'extraire l'ensemble des mots clés appartenant à un document. Cette extraction est effectuée en tenant compte des espaces, des chiffres et des ponctuations. C'est une étape qui peut sembler triviale au premier abord, et qui pourtant constituera la base de tout le reste du processus d'indexation.

2.1.2 Elimination des mots vides

Les mots vides sont des mots trop fréquents peu significatifs et porteurs de peu de sens, Le fait de ne pas éliminer les mots vides provoque inévitablement du bruit. Les mots vides représentent un facteur qui a une grande influence sur la précision de la recherche.

2.1.3 Normalisation des mots clés

Ce traitement consiste à retrouver pour un mot sa forme normalisée (généralement le masculin pour les noms, l'infinitif pour les verbes, le masculin singulier pour les adjectifs, etc.). Cette phase peut également être enrichie avec un traitement syntaxique et sémantique des mots-clés.

2.1.4 Pondération

Cette étape est entièrement dépendante du modèle de recherche d'information utilisé. Vu que les termes d'un document, n'ont pas tous la même importance, la pondération permet de définir le poids qu'a un terme dans un document donné.

Parmi les approches de pondération existantes, il y a le TF-IDF.

- **Pondération basée sur TF-IDF**

Le Tf,'Term frequency' ou fréquence d'un terme est l'ensemble d'occurrences de ce terme dans le document. Le tf est donné par plusieurs formules, citons par exemple :

$$tf = f(t,d) \Big/ \max([f(t,d)])$$

$$tf = \log(f(t,d))$$

$$tf = \log(f(t,d) + 1)$$

Où *f(t,d)* est la fréquence d'occurrence du terme ***t*** dans le document ***d***;

L'IDF, inverse document frequency est une mesure de l'importance du terme dans le corpus. Elle vise à donner plus de poids aux termes moins fréquents considérés plus discriminants **[Jaillet, 03]**. Il est donné par la formule suivante :

$$idf = N/n$$

Où ***N*** est le nombre de documents dans le corpus, et ***n*** ceux qui contiennent le terme ***t***.

Une formule tf-idf peut être la multiplication d'une tf par une idf.

3. Modèles de la recherche d'information

Un modèle de la RI doit fournir une formalisation du processus de recherche d'information et accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence **[Sauvagnat, 05].**

On retrouve dans la littérature plusieurs modèles de recherche, le tableau ci-dessous illustre une classification de ces modèles.

Modèles ensemblistes	Modèles algébriques	Modèles probabilistes
- Modèle booléen	- Modèle vectoriel	- Binary Independence Model
- Modèle booléen étendu	- Modèle vectoriel généralisé	- Modèle inferentiel bayésien
- Modèle booléen flou	- Latent Semantic Indexing	- Modèle de langage

Tableau02 : *Classification des modèles de la RI*

3.1 Le modèle vectoriel

Le modèle vectoriel est basé sur des statistiques qui ont pour but d'une part de caractériser d'un point de vue quantitatif les termes et les documents et d'autre part de mesurer le degré de pertinence d'un document vis à vis d'une requête. La pertinence d'un document par rapport à une requête revient à calculer la mesure de similarité vectorielle entre les deux vecteurs $\overrightarrow{d}$ (document) et $\overrightarrow{q}$ (la requête). Il existe plusieurs mesures de similarité pour ce modèle, la plus simple est le produit scalaire :

$$produit\ scalaire\ (\vec{d},\vec{q}) = \vec{d}.\vec{q} = \sum_{i=1}^{n} w_{i,d} * w_{i,q}$$

La figure Fig18 illustre un exemple d'une représentation vectorielle.

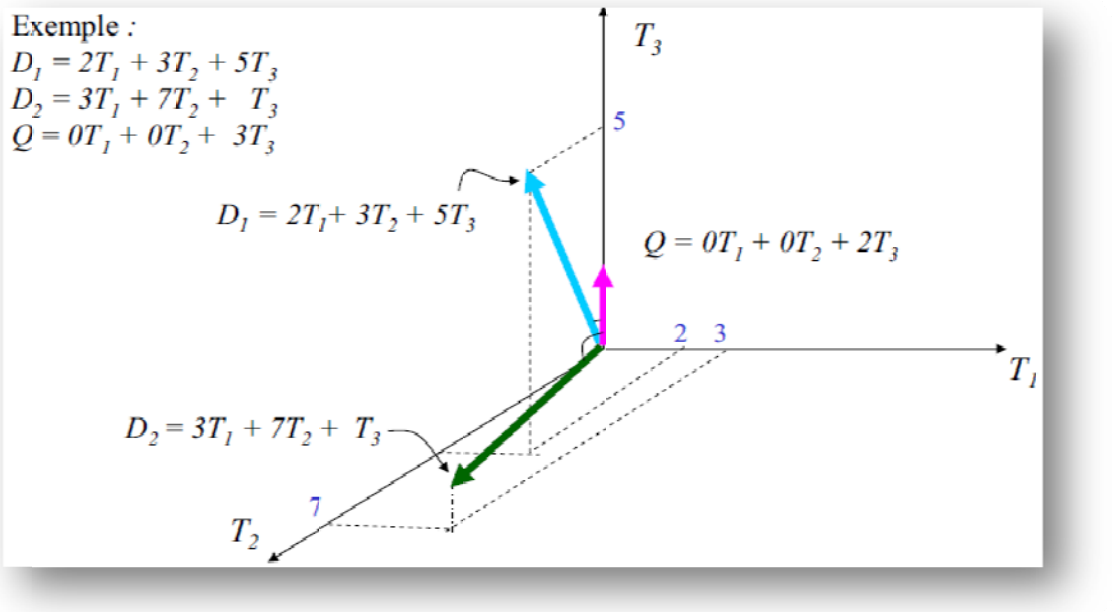

Fig 18 : *Exemple de représentation vectorielle*