Le web social et le web sémantique pour la recommandation de ressources pédagogiques

Mérième Ghenname




ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE ET D'ANALYSE DES SYSTÈMES - UNIVERSITÉ MOHAMMED V RABAT

&

TÉLÉCOM SAINT-ETIENNE - UNIVERSITÉ JEAN MONNET SAINT ETIENNE

T H È S E DE DOCTORAT

pour obtenir le grade de Docteur en informatique

LE WEB SOCIAL ET LE WEB SÉMANTIQUE POUR LA RECOMMANDATION DE RESSOURCES PÉDAGOGIQUES

Présentée et soutenue publiquement par

MÉRIÈME GHENNAME

Le 02 décembre 2015

Directeurs de thèse : **Rachida AJHOUN, Frédérique LAFOREST**

Encadrants de thèse : **Mounia ABIK, Christophe GRAVIER, Julien SUBERCAZE**

Formation doctorale : Informatique

MEMBRES du JURY :

M. Rochdi MESSOUSSI Université Ibn Tofail, Kénitra	Professeur	(***Président***)
Mme. Rachida AJHOUN Université Mohammed V de Rabat	Professeur	(***Directeur***)
Mme. Frédérique LAFOREST Université Jean Monnet Saint-Etienne	Professeur	(***Directeur***)
Mme. Mounia ABIK Université Mohammed V de Rabat	Professeur Habilité	(***co-Directeur***)
M. Christophe GRAVIER Université Jean Monnet Saint-Etienne	Maître de conférence	(***co-Directeur***)
M. Julien SUBERCAZE Université Jean Monnet Saint-Etienne	Chargé de recherche	(***invité***)
M. Karim BAÏNA Université Mohammed V de Rabat	Professeur Habilité	(***Rapporteur***)
M. Omar EL BEQQALI Université Sidi Mohamed BenAbdellah, Fès	Professeur	(***Rapporteur***)
M. Serge GARLATTI Institut Mines Telecom, Telecom Bretagne	Professeur	(***Rapporteur***)

Remerciements

Soyons reconnaissants aux personnes qui nous donnent du bonheur ; elles sont les charmants jardiniers par qui nos âmes sont fleuries, "Marcel Proust". Ce travail de thèse de doctorat n'aurait jamais abouti sans l'aide, les encouragements et l'implication de certaines personnes à qui j'exprime à travers ces quelques phrases modestes toute ma gratitude.

Tous mes remerciements vont vers mes deux directeurs de thèse Mme Rachida Ajhoun professeur du côté de l'université Mohamed V de Rabat et Mme Frédérique Laforest professeur du côté de l'université Jean Monnet Saint-Etienne, de m'avoir accueilli et de m'avoir permis d'effectuer ce travail de mémoire dans de bonnes conditions. Je remercie également mes co-encadrants Mme Mounia Abik du côté de l'université Mohamed V de Rabat, et Mr Christophe Gravier et Mr Julien Subercaze du côté de l'université Jean Monnet Saint-Etienne. Je vous remercie tous de m'avoir donné toutes ces idées, ainsi que pour toutes les riches réunions d'encadrement, sans quoi je ne serais jamais arrivée à bout de ce travail, je ne vous remercierais jamais assez pour m'avoir fait profiter de votre expérience et de m'avoir témoigné tant de bienveillance. Merci pour votre compétence, votre rigueur scientifique et votre clairvoyance qui m'ont beaucoup appris. Je remercie chaleureusement tous les membres de jury :

- Monsieur Rochdi MESSOUSSI, professeur à L'Université Ibn Tofail, Kénitra de m'avoir fait l'honneur de présider le jury de cette thèse.
- Monsieur Omar EL BEQQALI, professeur à l'Université Sidi Mohammed BenAbdellah de Fès, Monsieur Serge GARLATTI, professeur à l'Institut Mines Telecom, Telecom Bretagne et Mr Karim BAÏNA professeur à l'Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes d'avoir accepté de juger ce mémoire.

Je tiens également à remercier tous les membres de l'équipe LeRMA de L'Ecole Nationale Supérieure de l'Informatique de d'Analyse des systèmes, et L'équipe Satin du Laboratoire Hubert Curien de Télécom Saint-Etienne pour leur ouverture et grand esprit de partage et d'échange de connaissances.

Je tiens aussi à exprimer ma gratitude à mes parents qui ont consacré leur vie pour moi, ma petite famille mes deux soeurs en témoignage des liens solides qui nous unissent, en vous souhaitant un avenir plein de succès et de bonheur. J'exprime mes fidèles sentiments de respect et de fraternité à mes deux meilleurs amis qui m'ont soutenu jusqu'au bout Nada Maaroufi et Youssef Baddi. Mes plus vifs remerciements

sont destinés à mon âme soeur pour son soutien inconditionnel, sa compréhension, sa disponibilité et sa grande confiance à mon cher Yassine.

Enfin, à tous ceux que je ne nomme pas, mais qui se reconnaîtront. En témoignage de mon sincère dévouement de mes reconnaissances pour vos sacrifices et vos encouragements, veuillez trouver dans ce modeste travail le fruit de toutes vos peines et vos efforts.

Le Web Social et le Web Sémantique pour la Recommandation de Ressources Pédagogique

Résumé :

Ce travail de recherche est conjointement effectué dans le cadre d'une cotutelle entre deux universités : en France l'Université Jean Monnet de Saint-Etienne, laboratoire Hubert Curien sous la supervision de Mme Frédérique Laforest, M. Christophe Gravier et M. Julien Subercaze, et au Maroc l'Université Mohamed V de Rabat, équipe LeRMA sous la supervision de Mme Rachida Ajhoun et Mme Mounia Abik. Les connaissances et les apprentissages sont des préoccupations majeures dans la société d'aujourd'hui. Les technologies de l'apprentissage humain visent à promouvoir, stimuler, soutenir et valider le processus d'apprentissage. Notre approche explore les opportunités soulevées en faisant coopérer le Web Social et le Web sémantique pour le e-learning. Plus précisément, nous travaillons sur l'enrichissement des profils des apprenants en fonction de leurs activités sur le Web Social. Le Web social peut être une source d'information très importante à explorer, car il implique les utilisateurs dans le monde de l'information et leur donne la possibilité de participer à la construction et à la diffusion de connaissances. Nous nous focalisons sur le suivi des différents types de contributions, dans les activités de collaboration spontanée des apprenants sur les réseaux sociaux. Le profil de l'apprenant est non seulement basé sur la connaissance extraite de ses activités sur le système de e-learning, mais aussi de ses nombreuses activités sur les réseaux sociaux. En particulier, nous proposons une méthodologie pour exploiter les hashtags contenus dans les écrits des utilisateurs pour la génération automatique des intérêts des apprenants dans le but d'enrichir leurs profils. Cependant les hashtags nécessitent un certain traitement avant d'être source de connaissances sur les intérêts des utilisateurs. Nous avons défini une méthode pour identifier la sémantique de hashtags et les relations sémantiques entre les significations des différents hashtags. Par ailleurs, nous avons défini le concept de Folksionary, comme un dictionnaire de hashtags qui pour chaque hashtag regroupe ses définitions en unités de sens. Les hashtags enrichis en sémantique sont donc utilisés pour nourrir le profil de l'apprenant de manière à personnaliser les recommandations sur le matériel d'apprentissage. L'objectif est de construire une représentation sémantique des activités et des intérêts

des apprenants sur les réseaux sociaux afin d'enrichir leurs profils. Nous présentons également notre approche générale de recommandation multidimensionnelle dans un environnement d'e-learning. Nous avons conçu une approche fondée sur trois types de filtrage : le filtrage personnalisé à base du profil de l'apprenant, le filtrage social à partir des activités de l'apprenant sur les réseaux sociaux, et le filtrage local à partir des statistiques d'interaction de l'apprenant avec le système. Notre implémentation s'est focalisée sur la recommandation personnalisée.

Social Web and Semantic Web for Recommendation in e-learning

Abstract :

This work has been jointly supervised by U. Jean Monnet Saint Etienne, in the Hubert Curien Lab (Frederique Laforest, Christophe Gravier, Julien Subercaze) and U. Mohamed V Rabat, LeRMA ENSIAS (Rachida Ahjoun, Mounia Abik).

Knowledge, education and learning are major concerns in today's society. The technologies for human learning aim to promote, stimulate, support and validate the learning process. Our approach explores the opportunities raised by mixing the Social Web and the Semantic Web technologies for e-learning. More precisely, we work on discovering learners profiles from their activities on the social web. The Social Web can be a source of information, as it involves users in the information world and gives them the ability to participate in the construction and dissemination of knowledge. We focused our attention on tracking the different types of contributions, activities and conversations in learners spontaneous collaborative activities on social networks. The learner profile is not only based on the knowledge extracted from his/her activities on the e-learning system, but also from his/her many activities on social networks. We propose a methodology for exploiting hashtags contained in users' writings for the automatic generation of learner's semantic profiles. Hashtags require some processing before being source of knowledge on the user interests. We have defined a method to identify semantics of hashtags and semantic relationships between the meanings of different hashtags. By the way, we have defined the concept of Folksionary, as a hashtags dictionary that for each hashtag clusters its definitions into meanings. Semantized hashtags are thus used to feed the learner's profile so as to personnalize recommendations on learning material. The goal is to build a semantic representation of the activities and interests of learners on social networks in order to enrich their profiles. We also discuss our recommendation approach based on three types of filtering (personalized, social, and statistical interactions with the system). We focus on personalized recommendation of pedagogical resources to the learner according to his/her expectations and profile.

Mots clés / **Key words :** Hashtags ; Social Network ; Natural Language Processing ; Clustering ; Semantic Web ; E-learning environments ; Recommendation

Discipline : Informatique

[1] LeRMA ENSIAS, Université Mohammed V de Rabat. Avenue Mohamed Ben Abdellah Regragui, 10000, Rabat MOROCCO
[2] LT2C, Telecom Saint-Etienne, Université Jean Monnet.25 rue du docteur Remy Annino, 42000 Saint-Etienne FRANCE

Liste de publications

- **EDUCON 2012** M. Ghenname, R. Ajhoun, C. Gravier, & J. Subercaze. Combining the semantic and the social web for intelligent learning systems. In Global Engineering Education Conference (EDUCON), 2012 IEEE, pages 1-6. IEEE, 2012.
- **ISKO 2013** Ghenname, M., Abik, M., Ajhoun, R., Subercaze, J., Gravier, C., & Laforest, F. (2013, November). Personalized recommendation based hashtags on e-learning systems. In ISKO-Maghreb, 2013 3rd International Symposium (pp. 1-8). IEEE.
- **ICDEW 2014** Ghenname, M., Subercaze, J., Gravier, C., Laforest, F., Abik, M., & Ajhoun, R. (2014, March). A hashtags dictionary from crowdsourced definitions. In Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on (pp. 39-44). IEEE.
- **IRECOS 2015** Ghenname, M., Abik, M., Subercaze, J., Gravier, C., Laforest, F., & Ajhoun. Hashtag-based learning profile enrichment for personalized Recommendation in ELearning Environments. Journal International Review on Computers and Software (IRECOS). Volume 10, Issue 9, September 2015.

Table des matières

CE DOCUMENT A ÉTÉ ÉDITÉ EN PDFTEX, AVEC L'ÉDITEUR TEXMAKER 4.2. SOUS WINDOWS 8.1.

Table des figures

Liste des tableaux

Liste des abréviations

SN	Social Network
SNA	Social Network Analysis
OWL	Web Ontology Language
URI	Uniform Resource Identifier
XML	eXtensible Markup Languager
RDF	Resource Description Framework
W3C	World Wide Web Consortium
MCL	Markov Clustering
HAC	Hierarchical Ascendant Classification
LMS	Learning Management System
PAPI	Public and Private Information
IMS-LIP	IMS Learner Information Package
IMS-RDCEO	IMS Reusable Definition of Competency or Educational Objectives
IMS-LIP	IMS Learner Information Package
Moodle	Modular Objet-Oriented Dynamic Learning Environment

CHAPITRE 1

INTRODUCTION

Sommaire

1.1 Contexte et problématique scientifique

1.1.1 Contexte de la thèse

Cette thèse fait l'objet d'une coopération entre l'équipe Satin de Télécom Saint-Etienne en France et l'équipe LeRMA de l'ENSIAS à Rabat au Maroc.

Les environnements informatiques pour l'apprentissage à distance (e-learning) ont pour objectif de favoriser ou susciter des apprentissages, de les accompagner et de les valider. Ce terme, né dans les années 90, souligne la complexité de ces systèmes qui ont leur source à la fois dans les travaux sur l'apprentissage humain et dans les travaux sur la modélisation des connaissances et des interactions. Nos travaux s'inscrivent dans le champ de la modélisation des connaissances et des interactions, et plus particulièrement dans la mouvance du Web sémantique. L'environnement informatique de notre société a très rapidement évolué ces dernières années, avec l'imprégnation extraordinairement importante des réseaux sociaux en ligne, qui s'étend dans tous les domaines de la société. Les environnements d'apprentissage à distance n'y échapperont pas, leur évolution sous la forme de Mooc est en route.

Aujourd'hui nous ne sommes pas face à un problème de manque de ressources, car les plate-formes d'apprentissage contiennent un grand nombre de ressources et offrent même plusieurs propositions de cours aux apprenants en fonction de plusieurs critères à savoir les intérêts, le niveau, les préférences etc. Ces informations sont dans la plupart des cas renseignées par les apprenants lors de la première inscription sur la plateforme ou bien issues des interactions de l'apprenant avec le système. Toutefois, un grand nombre d'informations concernant l'apprenant ne sont pas incluses dans leur profil d'apprentissage sur la plate-forme alors qu'elles existent bien sur les réseaux sociaux. Par conséquent, des ressources intéressantes et disponibles sur la plate-forme de e-learning ne sont pas proposés aux apprenants en raison du manque d'informations sur leurs profils. Les réseaux sociaux apportent des moyens d'échanger des informations et des connaissances entre les membres de ces réseaux, que ces membres soient des hommes ou des femmes, ou qu'ils représentent des entités comme des entreprises, des universités ou des sociétés savantes. Des passerelles entre les systèmes d'e-learning actuels et les réseaux sociaux doivent être étudiées et mises en place. Les outils du Web social et du Web sémantique permettent d'imaginer des solutions pour une recommandation multidimensionnelle dans le cadre de ce travail de thèse.

Le travail au sein de l'équipe Satin est consacré à la définition des méthodes d'analyses des connaissances permettant l'enrichissement du profil utilisateur, tandis qu'au LeRMA nous définissons les méthodes d'exploitation des connaissances extraites des réseaux sociaux pour faire des recommandations de ressources pédagogiques à l'apprenant lors de son apprentissage sur une plateforme e-learning.

1.1.2 Motivations et axes de recherche

Partant du fait que les étudiants inscrits sur les différents réseaux d'apprentissage de l'école (Moodle, Claroline, etc.), le sont majoritairement aussi sur des réseaux sociaux (RS) (facebook, twitter, google+, etc.), et que les prérogatives de l'apprentissage sur les RS dépassent ceux de la classe traditionnelle, il serait judicieux d'exploiter les activités des apprenants sur les RS dans le cadre formel. Notre idée consiste à analyser les traces et données partagées par les apprenants sur les RS dans le but de les réutiliser au sein du LMS. En d'autres termes il s'agit d'établir des passerelles de communications entre les réseaux sociaux et l'environnement d'apprentissage de l'apprenant. Nous avons fait le tour d'horizon des travaux de recherche orientés dans le

sens de la sémantique appliquée au Web qui est un domaine très vaste comprenant un grand nombre d'acteurs. Certains oeuvrent pour le web sémantique dont le but est de définir des méthodes pour rendre les données plus intelligibles et connectées [4, 5]. D'autres travaillent sur des projets concrets d'enrichissement de données [6, 7, 8]. Et d'autres exploitent les données sémantiques déjà existantes pour améliorer leurs services (comme le graphe social par exemple) [9, 10, 11]. L'enrichissement sémantique est une discipline très abordée actuellement, son potentiel de développement est énorme et ses applications infinies.

L'objectif de ce travail est d'utiliser les informations des adhérents des réseaux sociaux en tirant profit des nouvelles technologies du Web Sémantique pour améliorer la qualité de l'apprentissage. Il s'agira d'enrichir le profil de l'apprenant à partir des connaissances sur ses activités et de proposer une approche de recommandation en fonction de ce profil enrichi. Ce travail contient donc deux volets : un premier volet consacré à la définition des méthodes d'analyse des connaissances permettant l'enrichissement du profil, et un deuxième volet dédié aux méthodes d'exploitation des connaissances pour la recommandation des ressources pédagogiques dans un environnement d'e-learning.

Méthode d'analyse des connaissances pour l'enrichissement du profil Cette partie vise à rendre exploitables les traces laissées par les apprenants sur les réseaux sociaux. Dans ce travail nous construisons une représentation sémantique des thèmes traités dans les conversations de l'apprenant au sein des réseaux sociaux. Nous nous appuyons principalement sur les hashtags, et leur appliquons des traitements visant à expliciter leur sémantique. Ils deviennent alors exploitables pour l'enrichissement du profil de l'apprenant.

Méthode de recommandation des ressources pédagogiques à l'apprenant Le but de cette partie est de maximiser l'efficacité du processus d'apprentissage en e-Learning, en prenant en considération le profil de l'apprenant enrichi avec ses intérêts déduits de ses activités sociales. Nous nous attaquons dans cette partie à la modélisation de l'apprenant et à notre approche de recommandation multidimensionnelle dans un environnement d'e-learning.

1.1.3 Contributions

Visions autour de la convergence entre le Web social et le Web sémantique

Dans nos travaux nous nous intéressons à l'utilisation des technologies sémantiques afin de faciliter la réutilisation de l'information présente sur le Web social, et ce de manière utile et enrichissante pour un apprenant à la fois actif sur des réseaux sociaux et inscrit sur une plate-forme d'apprentissage d'e-learning. Les travaux que nous présentons dans cette thèse ont pour objectif d'aider les plates-formes d'e-learning à mieux cerner les besoins d'un apprenant en ayant plus d'informations sur ses intérêts. Ses intérêts résident dans les hashtags utilisées dans ses écrits sur les réseaux sociaux. Cet objectif est décliné par la construction d'un dictionnaire désambiguïsant les hashtags les plus utilisés au sein des réseaux sociaux, un système d'enrichissement du profil de l'apprenant en fonction des hashtags utilisés dans ses écrits et ses données partagées au sein de ses réseaux sociaux, et un système de recommandation personnalisée de ressources pédagogiques au sein d'une plate-forme d'e-learning à base du profil enrichi de hashtags transformés en intérêts.

Modélisation et formalisation

En termes de modélisation, nous avons opté pour un modèle structuré avec la spécification IMS Learner Information Package (IMS-LIP) pour la modélisation du profil d'apprentissage d'un apprenant, étant donnée sa capacité à faciliter l'import, et l'export des collections de données entre les systèmes éducatifs. Nous avons également choisi de modéliser nos hashtags et les définitions recueillies des sources en ligne selon le modèle de base de données NoSql Cassandra. Cassandra permet d'avoir des schémas de données flexibles grâce à sa représentation en colonnes, et elle est très rapide pour manipuler un volume important de données. De plus son architecture lui permet d'évoluer sans problème dans un environnement distribué.

Pour faire une abstraction de la sémantique des différents concepts abordés dans cette thèse nous avons aussi eu recours à une formalisation mathématique. Le problème à formaliser est précisément : **"la mesure de similarité entre les définitions d'un hashtag afin de les regrouper en unités de sens et ainsi générer le folksionary"**. Nous avons formaliser les concepts suivants : **simMeaning** pour la similarité entre deux définitions d'un hashtag, la fonction $\mathcal{E}$ qui relie chaque définition $def \in D(w)$ d'un hashtag donné à un sens $s \in S(w)$ du même hashtag et **Dist(w)** une matrice nor-

malisée qui exprime les distances entre les définitions d'un hashtag.

Réalisation logicielle

En plus des modélisations et formalisation mathématiques, nos travaux de thèse ont également abouti à des réalisations logicielles pour la validation de nos approches théoriques.

D'une part nous avons mis en place un ensemble de modules permettant de traitement des données sociales et la construction du folksionary. Ainsi nous avons utilisé le Web scraping jScraper pour le recensement des définitions d'un hashtag sur les dictionnaires en ligne et le classificateur de langue Apache Tika pour filtrer que les définitions en anglais. Nous avons réalisé une implémentation Java de l'algorithme Extended Lesk pour le calcul de distance sémantique entre les définitions d'un hashtag, et de l'algorithme Markov clustering pour le clustering des définitions d'un hashtag en unités de sens et finalement nous nous sommes basés sur l'algorithme de clustering hiérarchique CHA pour un clustering hiérarchique de sens de hashtags.

D'autre part nous avons implémenté des méthodes pour l'enrichissement du profil et la recommandation personnalisée. De ce fait nous avons fait une implémentation de l'algorithme Simplified Lesk pour la génération de la personomie d'un apprenant, et une implémentation Java pour notre algorithme de recommandation. Nous avons aussi eu recours au package java.net pour établir la communication entre notre folksionary et la plateforme d'e-learning moodle.

1.2 Organisation du mémoire

1.2.1 Plan du mémoire

Le Manuscrit est composé de cinq chapitres dont l'introduction et la conclusion.

Chapitre 1 : INTRODUCTION Ce premier chapitre est introductif, il présente le contexte du travail, les différentes motivations et les axes de recherche abordés tout au long du travail.

Chapitre 2 : ANALYSE DE LA CONVERGENCE DU WEB SOCIAL ET DU WEB SEMANTIQUE POUR LA RECOMMANDATION DANS LES SYSTEMES D'E-LEARNING Ce premier chapitre donne un état de l'art détaillé sur les différents concepts mise en oeuvre dans ce travail de recherche. Il présente les différentes notions du Web Social et du Web sémantique, les systèmes de recommandation et le e-learning nécessaires pour la compréhension du travail réalisé. Ce chapitre évoque d'une part le rôle des réseaux sociaux dans l'incitation des usagers à participer à la construction de l'information. D'autre part il éclaircit le rôle primordial du Web sémantique dans la structuration et la valorisation des données, principalement celles issues du Web social. Il considère également l'effet de la cohabitation des deux paradigmes du Web pour une approche à double dimension : (1) l'introduction des pratiques communautaires de la communication pour une meilleur implication de l'apprenant (dimension Web Social), et (2) la personnalisation de l'information (dimension Web Sémantique). Ainsi il traite brièvement la question de l'enrichissement du profil apprenant en fonction de ses différentes activités sur les réseaux sociaux, et comment les technologies du Web sémantique permettent d'analyser les données générées par les médias sociaux et les transformer en connaissance exploitable au sein d'un environnement d'apprentissage. Ce chapitre traite également les approches de recommandation et particulièrement dans le cadre du e-learning. Nous concluons par une prise de position déterminant notre problématique.

Chapitre 3 : METHODE D'ANALYSE DES CONNAISSANCES SOCIALES POUR L'ENRICHISSENT DU PROFIL DE L'APPRENANT Ce chapitre est consacré à détailler les méthodes d'analyse de données et traces laissées sur les réseaux sociaux pour l'enrichissement du profil. Nous avons défini un moyen de caractériser les informations produites dans les réseaux sociaux avec le peu d'éléments sémantiques qu'elles contiennent. Cela nous a amenés à nous pencher sur les hashtags, ces balises écrites par les personnes au sein de leurs messages. Ces hashtags sont auto-émergents (n'importe qui peut créer n'importe quel hashtag) et mal définis (on trouve des définitions manuelles répertoriées dans des dictionnaires en ligne, et parfois contradictoires). Nous avons donc voulu construire en première instance un dictionnaire consolidé de hashtags, qui, à la manière d'un dictionnaire classique, associe à chaque hashtag un ensemble de sens, chaque sens pouvant avoir plusieurs définitions. On appelle ce dictionnaire un folksionary. Les outils du Web sémantique et

les techniques de data mining ont été utilisés pour arriver à ces fins. Nous avons ensuite procédé à une validation de nos résultats via une analyse qualitative mesurant la distance entre le folksionary généré par notre approche automatique et une vérité de terrain établie manuellement. Et finalement nous avons effectué un clustering hiérarchique entre les différents sens des différents hashtags du Folksionary, afin de dégager les relations non explicites qui peuvent exister entre les hashtags.

Chapitre 4 : VERS UNE APPROCHE SOCIALE SEMANTIQUE POUR UNE RECOMMANDATION PERSONNALISEE DES CONTENUS PEDAGOGIQUES Dans ce chapitre il s'agit d'élaborer un profil de l'apprenant à partir des connaissances sur ses activités déjà recensées et analysées, dont le but est d'étudier la recommandation personnalisée des ressources pédagogiques sur une plate-forme d'e-learning. L'idée dans ce chapitre est d'expliquer comment on peut exploiter les ressources des réseaux sociaux au sein du LMS afin d'établir des passerelles de communication entre les réseaux sociaux et l'environnement d'apprentissage de l'apprenant. Nous expliquons notre approche de construction de la personomie de l'apprenant qui regroupe les hashtags utilisés dans ses écrits. Nous donnons ensuite le détail de l'algorithme de transformation de la personomie en un ensemble d'intérêts de l'apprenant, ainsi que la méthode d'enrichissement du profil apprenant modélisé selon le standard IMS-LIP. Plus loin dans le chapitre nous testons notre approche de recommandation à base du profil enrichi dans sur une plate-forme d'e-learning (moodle dans notre cas).

Chapitre 5 : CONCLUSION Enfin nous concluons ce manuscrit en fournissant un aperçu global sur les différents travaux réalisés lors de ce travail avec un regard critique, tout en proposant des perspectives pour la suite des travaux dans un futur proche.

1.2.2 Guide de lecture

Dans le but de faciliter la lecture et de permettre au lecteur d'accéder facilement aux différentes parties du mémoire, nous proposons un guide de lecture (Figure 1.1). La chapitre 1 introduit le travail et présente le contexte et les motivations de ce travail de recherche. Pour un aperçu sur les travaux précédents et les concepts utilisés dans ce travail, ainsi qu'une idée générale et synthétique de la problématique et de

nos travaux de recherches, nous conseillons de lire les chapitre 1 et 2 . Les deux chapitres qui suivront (chapitres 4 et 5) détaillerons avec précision les problématiques de recherche et les différentes solutions apportées. Le dernier chapitre donne un regard critique des travaux et évoquera les perspectives et les travaux futurs à réaliser.

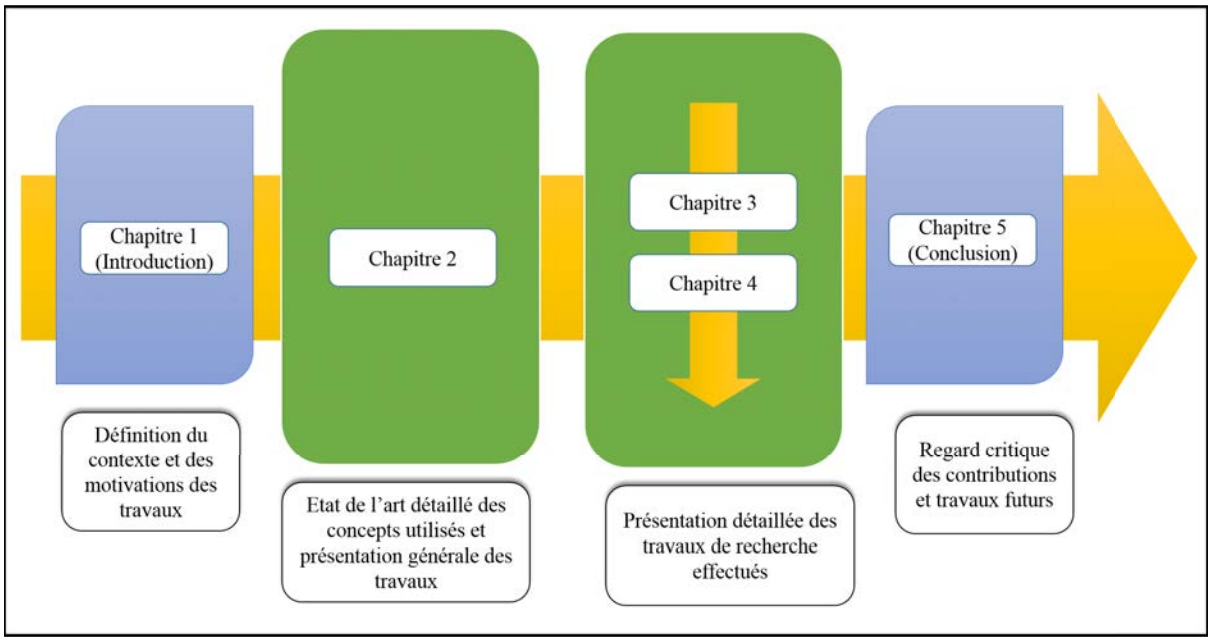

FIGURE *1.1: Organisation du mémoire*

CHAPITRE 2

ANALYSE DE LA CONVERGENCE DU WEB SOCIAL ET DU WEB SEMANTIQUE POUR LA RECOMMANDATION DANS L'E-LEARNING

Sommaire

Au cours de la dernière décennie, deux nouvelles facettes du Web ont émergé. D'une part, le Web sémantique fournit les moyens pour briser les barrières de communications entre les sources de données hétérogènes et distribuées présentes sur le Web. D'autre part, le Web social permet de regrouper les internautes en communautés, il met en avant les échanges, l'ouverture et la collaboration entre internautes par l'intermédiaire d'outils et de services simples d'utilisation. Les deux visions incluent de nombreuses applications en ligne, y compris l'éducation. Dans ce chapitre nous donnons un aperçu sur les travaux des pionniers et les possibilités soulevées par la cohabitation entre le Web Social et le Web Sémantique pour le e-learning.

2.1 Le Web social : participation des usagers dans le monde de l'information

Le Web n'a cessé d'évoluer depuis sa création, d'abord le Web 1.0 était une grande bibliothèque avec un grand nombre de documents où les utilisateurs pouvaient obtenir de l'information descendante. Le webmaster mettait en ligne des informations sur un site Web donné et les utilisateurs consultaient cette information sans possibilité d'interagir avec le site. L'information était à sens unique, jusqu'à ce que de nouveaux usages basés sur le partage de l'information soient venus l'enrichir.

Le terme Web 2.0 a été introduit par [12] comme résultat d'un remue-méninges au sein de l'entreprise de Tim O'Reilly, et le terme a été expliqué par ce dernier dans "What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software" [13]. Et ainsi se tiendra la première conférence Web 2.0 en octobre 2004, sous le thème "What Is Web 2.0", qui sera en suite publié en septembre 2005. Le Web 2.0 a complètement changé la vision du Web, en plaçant l'internaute au coeur de l'information, ainsi son rôle a évolué d'un simple récepteur passif d'information à un contributeur actif. C'est ce qui fait que le Web 2.0 est aussi synonyme du Web social, Web collaboratif, ou Web communautaire etc. Le Web 2.0 a changé la relation des usagers avec l'information. De plus son principe est de faire en sorte que l'information parvienne aux usagers plutôt qu'ils aillent la chercher. Le tableau 2.1 ci-après résume les principaux éléments de différences entre le Web 1.0 et le Web 2.0. [14].

Tableau 2.1: Eléments clés de l'évolution vers le Web 2.0 [14]

Web 1.0	**Web 2.0**
Fonctionnalités ou activités pour les usagers du Web	

- consultation - recherche dans des répertoires - achat - commentaire	- contribution au contenu - interaction par des moyens multiples - personnalisation des interfaces et des contenus - recherche dans les contenus
Outils disponibles pour l'usager	
- page web personnelle - formulaires à compléter - certains outils d'interaction (forums, évaluations, etc.)	- blogues - wikis - réseaux sociaux - agrégation et syndication - partage de médias
Objectifs et fonctionnalités pour les concepteurs de sites	
- diffusion large - mise en forme - hyperliens	- consultation - participation - mise à jour continues - applications composites (mashup)
Technologies sous-jacentes	

- pages statiques (essentiellement HTML)	- bases de données, XML, CSS, Ajax, RSS, etc.
Résumé en quelques mots	
- lecture seulement (read-only Web) - commerce	- lecture et écriture (read-write Web) - Web participatif - Communauté

Nous concluons que l'avancée principale du Web 2.0 par rapport au Web 1.0 c'est la transformation des utilisateurs de simples consommateurs des contenus à des membres actifs. Car grâce à ses nouvelles technologies le Web 2.0 permet de maximiser le potentiel de création et de partage et d'interactions des internautes avec les contenus du Web d'un coté, et de mettre en avant les échanges sociaux entre les utilisateurs d'un autre. Ces révolutions techniques et technologiques du Web 2.0 ont donné naissance à de nouveaux outils de communication dont les plus répondus sont les réseaux sociaux.

2.1.1 Web social et réseaux sociaux

Il n y a pas de définition stable et unique pour le Web 2.0. Mais le Web 2.0 peut être décrit comme l'ensemble des technologies et applications Web favorisant les échanges et rendant le Web un véritable espace de communication, tels que les sites de réseautages sociales, les outils de communication et les wikis etc. Son succès est principalement dû à l'émergence universelle de ses services tels que les blogs, les flux de syndication du contenu, et le marquage(ou tagging social) qui permettent aux utilisateurs de publier et partager du contenu facilement. Le terme Web 2.0 a été introduit par [12] en 2004. [12] définit le Web 2.0 à partir de cinq caractéristiques significatives [15] : (1) la première concerne la capacité des utilisateurs à créer et

partager des contenus et qui constitue l'élément central du développement récent du Web. Ces contenus produits par les utilisateurs de façon spontanée (User Generated Content UGC) se sont multipliés rapidement et excessivement depuis 2005 grâce aux plates-formes collaboratives (blog personnels, diffusion planétaire de photos, vidéo d'amateurs, réseaux sociaux etc.). (2) La deuxième révèle les nouvelles plateformes accessibles et facile à utiliser. (3) La troisième est le mode de collaboration entre les usagers et la création de communautés numériques autour de sujets d'intérêts. (4) La quatrième concerne la mise en valeur et la syndication des contenus et contributions des utilisateurs. (5) La cinquième et dernière est le fait que le Web social s'appuie sur la popularité des pratiques et usages qu'il est possible de rassembler en six catégories :

- **Les blogs :** ils font référence à un type de site utilisé pour la publication périodique et régulière d'articles, généralement succincts, et rendant compte d'une actualité autour d'un sujet donné ou d'une profession. A la manière d'un journal de bord, ces articles sont typiquement datés, signés et se succèdent dans un ordre antéchronologique, c'est-à-dire du plus récent au plus ancien [16].
- **Les activités d'échanges de fichiers :** tous les services d'échanges de fichiers vidéos, photos, musique (Flickr, Delicious, Picasa etc...) [15].
- **Les pratiques collaboratives d'écriture :** dont Wikipédia [1] constitue son visage emblématique [15].
- **Les sites de vente en ligne :** ils permettent aux usagers de poster leurs avis et commentaires à propos des produits(eBay Amazon...) [15].
- **Les univers virtuels appelés métaveres :** un métavère est un monde virtuel qui héberge une communauté d'utilisateurs présents sous forme d'avatars pouvant s'y déplacer, y interagir socialement et parfois économiquement. Ils peuvent également interagir avec des agents informatiques (Second Life, World of Warcraft...) [15].
- **Les sites de réseaux sociaux :**ils désignent l'ensemble des sites internet permettant de se constituer un réseau d'amis ou de connaissances professionnelles et fournissant à leurs membres des outils et interfaces d'interaction, de présentation et de communication [17].

Il est également important de faire la distinction entre les Réseaux Sociaux et le Web 2.0 qui représentent deux concepts différents souvent confondu. Ce qu'il faut

1. http://fr.wikipedia.org/wikiWikipédia:Accueil_principal

noté c'est que le Web 2.0 est à la fois une plate-forme sur laquelle les technologies innovantes ont été construites et un espace où les utilisateurs sont traités comme des objets [18]. Tandis que les réseaux sociaux désignent des sites dont la vocation première est la mise en relation des utilisateurs. Ils permettent de mettre en avant le partage des contenus, des opinions, des nouvelles, des expériences et des perspectives, mais surtout d'exister, d'influencer et d'accomplir des choses. Les réseaux sociaux ne constituent qu'une partie(certes non négligeable) du Web 2.0, sans pour autant le résumer.

Dans nos travaux, nous accordons une attention particulière aux réseaux sociaux et leurs apports pour le e-learning. Dans la suite du chapitre nous abordons plus en détail dans quel sens notre approche tire part des avantages des contributions des utilisateurs sur les réseaux sociaux pour améliorer l'efficacité des apprentissage, et nous citons également les différents travaux et contributions dans ce sens.

2.1.2 Gestion des connaissances par les utilisateurs

La tendance participative du Web 2.0 a permis de générer une masse données considérable qui ne cesse d'augmenter au fil du temps. La multitude d'informations instantanées et en continu a conduit vers l'infobésité (ou surcharge informationnelle), les utilisateurs reçoivent aujourd'hui 10 fois plus d'informations qu'il n'en recevait il y a 10 ans, en produisant environ 10% de plus chaque année et consacrent plus de 30% de leur temps de travail quotidien à cette activité, proportion qui ne cesse de croître ces cinq dernières années [19]. Face à cette abondance informationnelle un nouveau concept a jailli permettant aux utilisateurs d'accéder facilement à l'information, via l'indexation collaborative des ressources ou ce qu'on appelle les folksonomies qui sont considérées comme partie intégrante du Web 2.0.

2.1.2.1 Folksonomies et hashtags

Le hashtag est un terme qui revient très souvent dans les folksonomies, il désigne un mot-clé, une catégorie ou une métadonnée. Le mot hashtag signifiant en français : étiquette de balisage, étiquetage, fléchage, marquage, voire traçage, ou tagage collaboratif. Guy Marieke et Emma Tonkin [20] disent qu'une définition simple des hashtags serait la suivante :

Définition 1 *[20] : Les hashtags sont des mots-clés, des catégories de noms, ou des métadonnées. Essentiellement, un hashtag est simplement un jeu de mots-clés librement choisi. Les hashtags ne sont pas crées par des spécialistes de l'information, ils ne suivent aucune indication formelle. Cela signifie que ces items peuvent être catégorisés avec n'importe quel mot définissant une relation entre la ressource en ligne et un concept issu de l'esprit de l'usager. 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 hashtag.*

Le hashtag peut prendre toutes les formes possibles, selon le désir de l'utilisateur et surtout selon sa culture et sa maîtrise de la langue. Le tagging social ne repose sur aucun thésaurus et ne suit aucun modèle particulier, les hashtags peuvent être des mots absents dans le dictionnaire ou des néologismes. Le hashtag sert également à tisser des liens entre leurs utilisateurs. Souvent, il permet de mieux connaître les goûts ou les petites habitudes de leurs auteurs. Les hashtags sont généralement en accord avec le thème des publications qu'ils étiquettent. Bien que l'utilisation de hashtags possède beaucoup d'avantages, comme la souplesse et la facilité, ils engendrent des confusions et des polysémies. Un même hashtag peut avoir différents sens car il peut être lié à un ensemble de ressources.

L'utilisation de hashtags est liée à la pratique d'étiquetage ou de tagging, où un utilisateur associe un hashtag à une ressource donnée (billet de blog, photo ...). Cette relation qui forme ainsi une relation tripartite [21] peut se représenter par :

Tagging (Utilisateur, Ressource, Hashtag)

Telle que [21] :

- Utilisateur correspond à l'utilisateur qui effectue l'action ;
- Ressource correspond à la ressource annotée(billet de blog, pageWeb...) ;
- Hashtag correspond à l'étiquette attribué par l'utilisateur pour annoter une ressource ;
- Tagging correspond à l'action liant ces trois éléments.

Etant donné que plusieurs hashtags peuvent être associés par un même utilisateur à une même ressource, et qu'un même hashtag peut être associé à une même ressource par différents utilisateurs, et aussi une même ressource peut être associée à différents hashtags par différents utilisateurs. Les actions de tagging ne sont en général pas isolées, ce qui justifie l'appellation social tagging. Ainsi, la figure 2.1 représente

trois actions de tagging (T1, T2, T3) associés à une même ressource (photo) via deux utilisateurs (U1, U2) et deux tags distincts (mac, laptop) de la manière suivante :

- T1 (U1, mac, photo) ;
- T2 (U2, mac, photo) ;
- T3 (U2, laptop, photo).

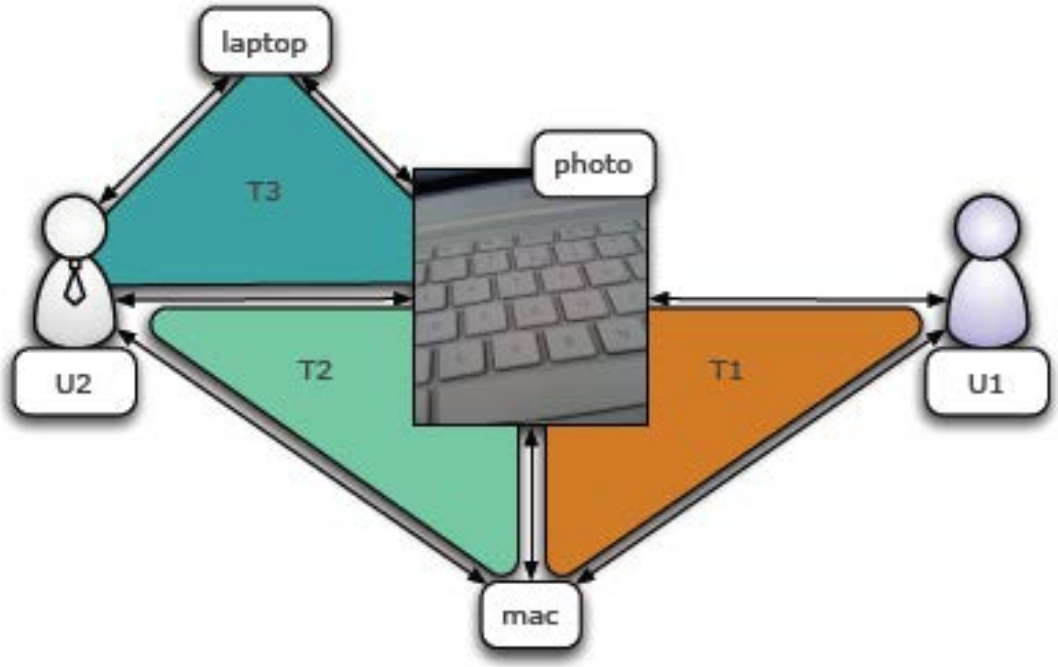

FIGURE *2.1: Actions de tagging combinées autour d'une même photo [1]*

2.1.2.2 Folksonomies

L'ensemble des hashtags constitue ce qu'on appelle une folksonomie [22], elle représente un moyen permettant à l'usager d'indexer des documents et de les retrouver facilement grâce à un système de mots-clés. Ce nouveau concept a surgit suite à l'accélération de la production d'informations et il diffère nettement des systèmes classificatoires classiques, comme la classification décimale universelle [23] ou la classification Dewey [24], qui s'inscrivent dans des processus plus longs et dont le but est d'obtenir un classement cohérent de documents physiques dont le contenu est inscrit dans la durée. Les folksonomies, contrairement aux autres systèmes ne reposent sur aucun thésaurus, ce qui donne à l'utilisateur une liberté totale quant au choix des

mots-clés. Les folksonomies sont donc centrées sur l'usager, ce dernier les utilise dans un but personnel, afin d'organiser son propre système d'information. Chaque usager s'organise en fonction de ses propres besoins en terme d'information, il n'est pas contraint d'une organisation pré-établie, ni d'un modèle qu'il faut suivre. Tout peut alors se trouver indexé par l'usager, depuis ses favoris jusqu'à ses photos, en passant par ses messages sur son blog. L'esprit collaboratif permet également d'indexer des documents produits par les autres en fonction de ses besoins.

Définition 2 *[22] Une folksonomie est une adaptation française de l'anglais folksonomy, mot-valise combinant les mots folk (le peuple, les gens) et taxonomy (la taxinomie). C'est un système de classification collaborative décentralisée spontanée, basé sur une indexation effectuée par des non-spécialistes. À l'inverse des systèmes hiérarchiques de classification, les contributeurs d'une folksonomie ne sont pas contraints à une terminologie prédéfinie mais peuvent adopter les termes qu'ils souhaitent pour classifier leurs ressources. Et ainsi une folksonomie est issue d'un ensemble d'actions de tagging et peut se formaliser comme suit.*

$$Folksonomie\ (Utilisateur^{+}, Resource^{+}, Hashtag^{+}, Tagging^{+})$$

Tel que :

- $Utilisateur^{+}$ *: correspond à un ensemble (fini) d'utilisateurs.*
- $Resource^{+}$ *: correspond à un ensemble (fini) de ressources annotées.*
- $Hashtag^{+}$ *: correspond à un ensemble (fini) de tags.*
- $Tagging^{+}$ *: correspond à l'ensemble de relations qui permettent de lier les éléments de ces différents ensembles, tel que défini précédemment.*

Thomas Vander Wal [25] distingue deux types de folksonomies :

- **Les folksonomies étroites** (narrow folksonomies) : utilisées à but personnel comme le système Flickr [2]. Dans ce type de folksonomies l'utilisateur indexe ses propres ressources uniquement.
- **Les folksonomies larges** (broad folksonomies) : utilisées à but collaboratif. A titre d'exemple les sites de partages de favoris, comme del.icio.us [3] ou Connotea [4] etc. Ce type de folksonomie permet à des utilisateurs d'être mis en relation suivant leurs centres d'intérêts, car non seulement il attribue des hashtags à ces propres ressources mais également à celles d'autrui.

2. https://blog.flickr.net/fr
3. https://delicious.com
4. http://www.connotea.org/

La force des folksonomies demeure dans le fait qu'elles ne nécessitent aucun consensus, à l'inverse des taxinomies. Il n'y a pas de politique dans les folksonomies. Il y a seulement l'acte de taguer et le résultat cumulé et amplifié de ces hashtags. L'intérêt des folksonomies est également lié à l'effet communautaire : pour une ressource donnée sa classification est l'union des classifications de cette ressource par les différents contributeurs. Ainsi, partant d'une ressource, et suivant de proche en proche les terminologies des autres contributeurs il est possible d'explorer et de découvrir des ressources connexes. Plusieurs sites fonctionnent exclusivement avec ce système de classification comme :

- Del.icio.us, site de partage de signets ;
- Flickr, site de partage de photos ;
- Wikipedia [5], encyclopédie participative ;
- Yoolink [6] qui permet de partager avec son réseau d'amis des liens Internet ;
- Knowledge Plaza [7], plateforme collaborative de gestion des connaissances.

Cependant malgré tout l'intérêt et la convivialité des folksonomies, elles ne sont pas exemptes de défauts. On notera principalement le bruit qu'elles génèrent, notamment à cause des homonymes, chose qui complique l'accès à l'information. Elles peuvent ainsi être source d'info-pollution, pour reprendre l'expression d'Eric Sutter [26]. Le fait que les folksonomies ne reposent sur aucun thésaurus fiable introduit de nombreuses confusions sémantiques et des polysémies dérangeantes, qui peuvent aboutir à des résultats de recherche inefficace et à une grande perte de temps. Le problème majeur avec l'indexation collaborative c'est qu'elle n'est pas exhaustive vu que les individus ont l'habitude de décrire les sources de manière subjective, en fonction de leurs intérêts, une autre limitation qui interfère aussi dans la qualité de l'indexation est la façon d'orthographier les hashtags, qui impacte l'exactitude et la pertinence des termes proposés et appréhende de plus en plus la question de l'ambiguïté.

De nombreuses recherches aujourd'hui ont pour objectif la désambiguisation des termes [27, 28]. Les Folksonomies ne peuvent pas satisfaire efficacement tous les besoins en termes de normalisation, de vocabulaire structuré, de relations sémantiques et hiérarchiques d'où la nécessité d'interagir avec d'autres approches de classification afin de bénéficier de leurs atouts et produire une indexation de qualité.

5. http://fr.wikipedia.org
6. http://www.yoolink.fr
7. http://www.knowledgeplaza.net

Dans ce qui précède nous avons abordé les différentes technologies et caractéristiques émanant du Web 2.0, dans la section suivante nous abordons l'influence de ces dernières sur l'apprentissage à distance.

2.1.3 Caractéristiques du Web 2.0 : quels impacts sur l'apprentissage à distance ?

Loin des caractéristiques technologiques du Web 2.0, les phénomènes plus sociaux se rapportant au Web 2.0 représentent des opportunités et des défis en formation à distance. Ces caractéristiques, très inter-reliées, comprennent notamment [14] :

- **La participation :** Le Web 2.0 est devenu à double sens [29], les utilisateurs créent quotidiennement du contenu et les sites alimentés par les internautes deviennent de plus en plus nombreux. La participation est donc l'élément central du Web 2.0 et cette participation massive change profondément la relation avec les contenus, les modes de communication. Ceci a favorisé en conséquence la collaboration et l'intelligence collective qu'O'Reilly résume dans sa citation : « The most important element of Web 2.0 is the complete digital democracy of ideas » [30].
- **La connectivité :** Les environnements Web 2.0 sont construits pour permettre l'interaction entre personnes et l'interconnexion entre les données, favorisent la création des liens et de communautés d'intérêts. C'est aussi bénéfique en éducation car la connectivité facilite l'intégration et l'enrichissement du savoir.
- **L'ouverture :** Comme le Web 2.0 est construit sur une logique de partage, il favorise la culture d'ouverture et de réutilisation. D'une part il est soutenu par des logiciels open source qui se sont multipliés dans de nombreux domaines, dont l'éducation avec des initiatives comme Open Academic [8]. Et d'autre part cette ouverture s'est aussi étendue aux contenus avec la croissance des dépôts de ressources libres comme Wikimedia Commons ou le Open Educational Resources Movement [9], car la diffusion large des contributions demeure un des intérêts de la participation en ligne.
- **La mobilité :** L'inconvénient de la distance s'est déjà énormément réduit depuis les années 90 avec la possibilité d'accéder aux mêmes ressources au même

8. http://openacademic.org/
9. http://www.oercommons.org/

moment et quel que soit la localisation des personnes. Avec le Web 2.0 les localisations ne sont plus fixes comme auparavant, la mobilité est prise en compte et les applications en ligne permettent de transporter ses contenus avec soi grâce aux dispositifs intelligents. Par ailleurs, des individus même très éloignés dans l'espace peuvent entretenir en ligne des relations continues. L'apprentissage devient nomade et planétaire.

- **L'instantanéité :** les modes de communication en Web 2.0 combinent la flexibilité que permet la communication asynchrone avec le caractère convivial et instantané du synchrone. Ainsi il a la spontanéité des échanges synchrones mais grâce aux possibilités de rediffusion et d'indexation, l'utilisateur peut avoir accès aux ressources en tout temps et en tout lieu.
- **La pérennité :** Toutes les informations numérisées partagées, et indexées, en plus d'être créées pour une diffusion instantanée, demeurent accessibles et s'ajoutent à l'identité numérique de chacun.
- **Le multimédia :** Facilité par la multiplication des équipements personnels de production numérisée (téléphones cellulaires, caméras numériques, caméras web, etc.), ainsi que de logiciels gratuits ou peu coûteux de traitement et de diffusion, l'univers textuel dans lequel nous avons évolué jusqu'ici devient rapidement multimédia avec des milliers de vidéos, fichiers audio ou photos etc. Ceci offre de nouvelles opportunités et facilités pour l'apprentissage.
- **La diversité :** Les utilisateurs communiquent par le moyen de nombreux outils en fonction de leurs préférences individuelles et de la nature de l'échange. Ils n'interviennent pas de la même façon, ne communiquent pas non plus le même type de contenu, même en utilisant des modes de communications similaires. Cette diversité dans le cas de l'éducation permet à chacun de choisir les moyens les plus appropriés pour mener à bien ses apprentissages.
- **L'abondance :** La quantité de données générées par les environnements Web 2.0 atteint des niveaux disproportionnés. Ces fleuves d'informations peuvent être un endroit où l'on peut trouver des données très pertinentes, mais aussi des labyrinthes où l'utilisateur est facilement submergé et se perd. Une information non seulement considérable mais également de qualités diverses. Le challenge est donc de trouver les moyens de générer et de cibler les informations appropriées aux besoins individuels de chaque usager.
- **La personnalisation :** Malgré l'aspect collaboratif du Web 2.0, il évolue vers

une grande personnalisation. Ceci s'applique sur les contenus partagés, qui sont plus représentatifs du profil particulier d'un individu, ses intérêts et ses opinions, etc. Ainsi chacun peut choisir, combiner et accéder aux ressources selon ses intérêts et préférences.

Nous concluons que le Web social transforme le panorama d'apprentissage en fournissant de grandes possibilités pour l'apprentissage en général. L'inclusion des caractéristiques sociales du Web 2.0 dans le processus d'apprentissage d'e-learning contribue positivement à l'amélioration des systèmes d'e-learning pour répondre efficacement aux besoins spécifiques de la nouvelle génération. Ainsi l'apparition du e-Learning 2.0 défie les modèles éducatifs actuels, et transforme les pratiques d'enseignement en posant de nouvelles exigences aux processus d'apprentissage et l'élaboration de nouveaux outils qui contribuent à une meilleur éducation et de nouvelles opportunités pour les apprenants. Dans la suite de nos travaux nous nous intéressons principalement à la dimension de personnalisation. Nous cherchons à travers notre approche de profiter des atouts du Web Social pour une meilleur personnalisation lors de la recommandation des ressources pédagogiques aux apprenants.

En produisant des données, l'apprenant donne des informations précieuses pour plus de personnalisation dans les ressources que l'environnement d'apprentissage lui propose. L'exploitation et le traitement des données qu'il génère au sein des plates-formes collaboratives demeure une étape incontournable de la personnalisation. Dans la section suivante nous soulignons les avantages de de la collaboration et de l'intelligence collective au sein des communautés virtuelles dans un environnement d'apprentissage.

2.1.4 Rôle de la collaboration et de l'intelligence collective au sein des communautés virtuelles dans l'apprentissage

La salle de classe est une miniature de la société, une petite communauté réunie dans un contexte donné et pour une durée déterminée et ayant comme objectif commun l'apprentissage. En repensant la classe nous déduisons que le temps consacré pour une séance de cours est très réduit pour bien cerner une tâche complexe qui mêle à la fois la recherche d'information, la coopération, la collaboration etc. D'où l'intérêt d'intégrer de nouvelles dimensions dans les modes d'enseignement traditionnel, en ayant recours à de nouveaux outils maîtrisés par les apprenants, le plus souvent en

dehors du circuit d'enseignement. Le succès des réseaux sociaux est lié aux contenus rédigés par leurs utilisateurs. Ils envahissent la sphère du travail en partant de la vie privée, ce sont des espaces d'échange, d'interactivité et de liberté d'expression où les individus se sentent plus à l'aise pour s'exprimer. Ils favorisent le travail participatif et collaboratif et modifient également le rapport au temps, vu que le temps de la classe peut s'étendre au gré des échanges de messages. Profiter des atouts des réseaux sociaux pour l'apprentissage à distance peut être avantageux pour améliorer l'efficacité. Par ailleurs beaucoup de contributions et travaux ont évoqué la question de la situation d'apprentissage en communauté virtuelle, notamment celle de la valeur ajoutée du groupe dans le processus d'acquisition de savoirs et de développement de compétences.

Au sein d'une classe ou dans une communauté virtuelle, l'apprentissage ne se fait pas sans les autres. Le côtoiement divers participants apporte des points de vue différents sur une situation donnée, ce qui rend les discussions entre apprenants très enrichissantes. A titre d'exemple une communauté de pratique, réunit de nombreux spécialistes dans une thématique donnée, dont certains sont débutants et d'autres expérimentés : les questions des uns et les réponses des autres sont aussi l'occasion d'apprentissages. Ainsi les membres d'une communauté virtuelle apprennent des contributions et des expériences des autres membres, étant donné que chacun donne le meilleur de lui-même pour y exister et être reconnu, ceci rend les échanges plus efficaces et productifs. Le travail collaboratif est un travail effectué en commun par plusieurs personnes qui mutualisent leurs connaissances et leurs compétences, et coordonnent leurs actions pour obtenir un résultat dont ils sont collectivement responsables. Il représente un excellent moyen de confronter les points de vue et les discussions engagées entre les membres d'un groupe [31]. De plus, dans le cas d'une formation à distance, les échanges et les collaborations évite aux apprenants l'impression d'isolement qui conduit dans la plupart des cas à abandonner la formation. En effet, travailler au sein du groupe et se préoccuper de tenir ses engagements concourent à maintenir la motivation des apprenants.

Divers travaux ont tenté de mieux cerner les prérogatives de travailler avec et par autrui dans une démarche d'apprentissage et comment les interactions entre pairs impactent les compétences cognitives individuelles. Dans [32], [33] et [34] les chercheurs ont tenté de révéler le rapport existant entre le scénario pédagogique et la dynamique des interactions lors d'une situation de travail collaboratif sur des structures sociales

(forums de discussion, réseaux sociaux etc.). Ils se sont focalisés sur les apports de la dimension sociale intégrée à une démarche collaborative pour le processus cognitif et le l'acquisition de compétences. A travers leur approche ils ont prouvé l'utilité de cette intégration dans l'amélioration des capacités et du rendement des apprenants. D'autres réflexions [35] et [36] se sont orientées vers une des théories modernes de l'apprentissage supportées par le Web social : le socioconstructivisme [37]. D'où émerge l'hypothèse que pour développer l'autonomie chez un apprenant il est nécessaire qu'il apprenne à travailler avec les autres, voire même de construire ses apprentissages avec et par autrui. Afin qu'il soit en mesure de trouver un sens à ses apprentissages et qu'il puisse idéalement construire sa propre place dans un collectif où l'autonomie de chacun résulterait d'une interdépendance de tous.

Plus récemment nous avons également vu l'émergence de l'utilisation des réseaux sociaux pour l'apprentissage des langues étrangères. Plusieurs institutions (Université Stendhal Grenoble III [10], Université du Luxembourg [11] se sont intéressées à l'évolution des ressources proposées par les outils du Web 2.0 et les réseaux sociaux dans l'enseignement et l'apprentissage des langues [38]. Suite à cela plusieurs études de terrain ont été faites pour la mise en place de dispositifs de formation sous forme de réseaux sociaux éducatifs, comme par exemple Campus FLE Education de l'Université de Léon [12], ou Foreigners [13] et Echanges Campus Education [14] à l'université de Lille. Les objectifs de ces dispositifs sont d'une part favoriser la communication et les échanges linguistiques, pour une meilleur compréhension et production orales chez les apprenants. D'autres part essayer de nouveaux outils Web (réseaux sociaux, blogs, podcast , etc.)dans une situation d'apprentissage de langues étrangères.

Il y a eu également l'apparition d'un nouveau type d'espaces collaboratifs qui ont été adoptés au début dans plusieurs universités américaines et se sont ensuite répandus dans des université françaises et suisses, les MOOCs (Massively Open Online courses). Ces espaces offrent des forums qui donnent la possibilité de créer des communautés que ce soit pour les étudiants, les professeurs ou les assistants pédagogiques [39]. Les MOOCs mettent en oeuvre également des outils basés sur des médias numériques faisant intervenir le réel et le virtuel [40]. Parmi les MOOCs qui

10. http://www.u-grenoble3.fr/
11. http://wwwfr.uni.lu/
12. http://flenet.unileon.es/courstourdumonde/Leon/leon.htm
13. http://foreignerinlille.ning.com/
14. http://lewebpedagogique.com/campusfle/

ont eu beaucoup de succès en ligne nous citons Coursera [15] dont l'un de ses cours diffusés a dépassé cent mille étudiants inscrits [40].

L'intérêt des travaux et expériences cités précédemment, est de souligner les apports des dimensions communicatives des outils Web 2.0 dans la mise en pratique des pédagogies constructivistes renforcées par les potentiels des TICE dans des démarches apprenantes. Cependant les recherches ne sont pas limitées juste à la création de réseaux sociaux à vocation pédagogique ou à l'emploi du travail collaborative au sein de communautés virtuelles pour améliorer les rapports étudiants/ enseignants. D'autres méthodes ont également essayé d'exploiter le bruit et les traces laissées par les internautes sur les réseaux sociaux, que ce soit dans un cadre d'apprentissage ou de loisirs d'où le jaillissement des SNA "Social Network Analysis". Dans la section suivante nous traitons le pouvoir du Web social à pallier le manque de données en enrichissant les connaissances sur les profils des utilisateurs et en soutenant les systèmes d'e-learning.

2.1.5 Le Web social pour pallier le manque de données d'usage

Le Web social contribue amplement à regrouper les usagers du Web dans des communautés d'intérêts, et les rend prosommateurs de leurs propre expérience du Web. Le Web 2.0 se distingue du Web 1.0 par cette prise de contrôle de l'information par les utilisateurs. En d'autres termes les utilisateurs se sont approprié le Web et ceci a mis en avant la production des contenus à forte valeur ajoutée. Par conséquent il y a eu une augmentation exponentielle des quantités de données disponibles en ligne. Le e-Learning n'est pas resté à l'écart de ces changements, et face à ces informations massives les systèmes ne se contentent plus de la simple présentation des données aux apprenants. Ils ont mis l'apprenant au centre de leurs démarches et ainsi la personnalisation devient l'élément clé du e-learning.

Les apprenants aujourd'hui apprennent plus hors du cadre formel et même sans le réaliser, en lisant, écoutant ou même regardant ou tout simplement en parlant les uns les autres. Les internautes sont actifs sur les réseaux sociaux les plus populaires comme : facebook, twitter et flickr. Beaucoup d'entre eux participent aussi activement à des réseaux en ligne spécifiquement axés sur l'éducation [41]. Les prérogatives de l'apprentissage sur les réseaux sociaux (SN) dépassent ceux de la classe tradition-

15. https://fr.coursera.org/

nelle. Il faut profiter des atouts de l'environnement social et de cette masse importante de traces laissées par les apprenants car plus on a d'informations sur l'apprenant mieux on pourra le servir.

Une question qui pourrait se poser ici c'est pourquoi ne pas se contenter d'utiliser directement les réseaux sociaux comme des environnements d'apprentissage ? La réponse réside principalement dans la difficulté rencontrée lors de la création des groupes à vocation pédagogique au milieu des réseaux sociaux. Le mélange entre les éléments de la vie privé et les éléments d'apprentissage ainsi que la présence des publicités représentent une incompatibilité avec les exigences de concentration demandées lors d'une démarche apprenante. Les réseaux sociaux constituent donc une aide précieuse dans l'apprentissage, ils peuvent motiver les étudiants, développer la rétroaction collaborative mais ne peuvent remplacer les plates-formes d'apprentissages. Le partage est l'élément qui crée la force des réseaux sociaux et qu'on pourra détailler plus clairement dans les points suivants :

- L'immédiateté car les réseaux sociaux diffusent les informations en temps réel, et les nouveautés sont clairement mises en avant.
- Le sentiment de liberté d'expression, qui permet à chacun de s'adresser à ceux qu'il souhaite, quand il souhaite et avec la fréquence qu'il souhaite, ces points dynamisent les relations entre les apprenants, et sont convenables pour développer des synergies d'apprentissages.

Cependant la convivialité offerte par les réseaux sociaux n'exclut pas le taux de bruit énorme, au milieu duquel se perd l'information qui devient difficile à exploiter. Nous avons alors besoin d'une part de solutions LMS pour satisfaire les exigences complexes tel que la prestation, le suivi et la validation des objectifs visés par le cours, un cadre formel avec des objectifs d'apprentissage structurés, une durée d'apprentissage et un soutien fourni qui entraîne à une certification. Et d'autre part de bénéficier des points forts des réseaux sociaux favorisant la collaboration à travers les liens sociaux.

Comme déjà tracé dans les objectifs de la thèse, le but final de nos travaux est la recommandation personnalisée dans une plate-forme d'e-learning. Mais le problème majeur des systèmes de recommandation reste le démarrage à froid. Plus précisément le rapprochement entre le besoin spécifique de l'utilisateur et le contenu devient une tâche très délicate lorsqu'on a pas assez d'informations au préalable sur nouvel utilisateur d'un système donné. Les informations disponibles sur les plates-formes ne

sont pas suffisantes pour une personnalisation raffinée. D'où l'idée de réutiliser les ressources sur les réseaux sociaux au sein du LMS pour des passerelles de communication entre les réseaux sociaux et l'environnement d'apprentissage de l'apprenant. Notre objectif est de combler cette carence de données d'usage via l'extraction et l'analyse des hashtags contenus dans les écrits laissés par les apprenants sur les réseaux sociaux. Les hashtags expriment très bien les intérêts de leurs utilisateurs, ainsi ils nous permettront d'enrichir le profil de l'apprenant avec cette information.

2.1.6 Les réseaux sociaux et la recommandation de contenus

2.1.6.1 Analyse des réseaux sociaux (SNA) pour la recommandation

Les réseaux sociaux sont un espace d'échange des opinions des uns et des autres, où n'importe quelle information peut rapidement prendre des proportions inattendues. Ceci est lié au fait qu'un utilisateur agit fortement sur son groupe social, et que le groupe peut également avoir en retour des influences sur les choix, les orientations, les comportements, et les opinions des utilisateurs. L'analyse des réseaux sociaux (SNA pour Social Network Analysis) permet d'analyser les relations entre les acteurs d'un réseau social, déterminer les liens et ainsi faire des rapprochements entre les personnes à travers leurs tendances et centres d'intérêts. Par conséquent l'analyse de ces réseaux sociaux leurs structures, et la position de chaque individu s'avère un moyen efficace pour interpréter les comportements et les rôles des intervenants dans un réseau social [42]. En d'autres termes comprendre dans quel sens une structure contraint des comportements, et produit des interactions et actions réciproques entre les éléments qui la constituent.

Le but de nos travaux étant d'exploiter les données communautaires des usagers des réseaux sociaux pour réaliser des recommandations personnalisées nous avons fait le tour d'horizon des travaux qui contribuent dans ce sens. L'analyse de réseaux sociaux inspire depuis peu la recommandation de contenus. Les auteurs dans [43] proposent une méthode combinant des données sociales avec des données thématiques relatives aux articles publiés pour faire de la recommandation de collaboration scientifique. Dans [44] les auteurs extraient les logs d'utilisations sociales de contenus à savoir les lectures, les écritures, les commentaires etc, afin de les exploiter pour produire une carte sociale résumant "qui parle de quoi". Ainsi selon une méthode de classification ils obtiennent des groupes de contenus et les communautés d'utilisa-

teurs et ce en regroupant les contributeurs associés aux contenus de chaque groupe. La recommandation de contenus et de personnes est réalisée en interagissant avec la carte : un zoom sur une communauté indique les contenus ordonnés par pertinence, les personnes importantes du groupe, etc.

Comme déjà mentionné auparavant, le Web social contient une grande quantité de données sous forme de textes fournis par les utilisateurs. Sous forme de statuts, commentaires, hashtags, etc. D'autres travaux s'intéressent alors aux graphes d'annotation de ressources et à la recommandation de contenus. Comme dans [45] et [46] où les auteurs analysent les commentaires textuels des internautes concernant des films. Mais la limitation principale de ces approches est la complexité des technologies avancées (analyse textuelle, linguistique, modèle de connaissance, etc.). Par ailleurs les hashtags sont généralement choisis de façon informelle et personnelle par les utilisateurs. Les travaux [47] et [48] s'intéressent à la modélisation des hashtags autant que des intérêts des internautes et construisent des profils à partir des hashtags laissés sur divers sites. Mais la limitation de ces approches réside dans l'utilisation des hashtags sans prendre en considération l'aspect sémantique. Dans ce sens les auteurs de [49] proposent une méthode basée sur un graphe tripartite utilisateur-objet-tag. Ils présentent trois algorithmes permettant de calculer le degré d'intérêt d'un utilisateur à un objet à partir de ce graphe. Les résultats expérimentaux ont démontré que l'algorithme proposé donne des recommandations avec une grande précision. D'autres travaux plus récents ont également exploré la détection de communautés sur les graphes tripartites comme [50], [51], et [52].

Les travaux précédents sont orientés vers l'analyse des données collectées dans le contexte d'un site et visent à produire des recommandations aux internautes de ces mêmes sites. Une autre approche différente des précédentes proposée par [53] a pour but de faire ressortir des recommandations à destination d'utilisateurs "externes", i.e. d'apprendre, à partir des traces laissées par des internautes du Web, des connaissances destinées à être exploitées au sein d'un autre site. D'autres approches comme [54] et [55] proposent un modèle de recommandation qui se base sur le filtrage collaboratif des comportements, elles exploitent les observations relatives aux comportements de navigation des utilisateurs pour produire des recommandations. Ce modèle vise à améliorer la qualité des prédictions et de garantir la robustesse du système de recommandation.

Dans [56], [57], [58] et [59], les auteurs ont travaillé sur des améliorations du mo-

dèle de recommandation. Ils ont combiné le filtrage collaboratif comportemental avec une approche de clustering calculant les clusters selon les similarités de voisins entre utilisateurs. Ce modèle utilise les associations transitives et les méthodes de prédiction de liens afin d'établir de nouvelles relations entre les utilisateurs. Ce modèle a pour objectif de faire face au problème de manque de données, de réduire l'espace de recherche des voisins et d'améliorer le temps de calcul des recommandations ainsi que leur précision. Ensuite dans d'autres contributions [60] et [61] les auteurs ont étendu le modèle pour détecter des leaders dans l'objectif de remédier au problème de démarrage à froid dans le cadre d'un réseau comportemental. Ces leaders ont la particularité d"être au "centre" de ce réseau et disposent d'une potentialité importante de prédiction des appréciations des autres utilisateurs concernant les nouveaux items introduits dans le système.

2.1.6.2 Recommandation sociale basée sur la confiance

D'autres techniques de recommandation plus récentes sur les réseaux sociaux sont basées sur la confiance (trust-based recommendation). Ce type de recommandation sociale se divise en deux principales étapes : (1) construire un modèle de confiance et (2) se servir d'un modèle de calcul pour estimer le degré d'intérêt d'un utilisateur pour une ressource [62].

- **Le modèle de confiance :** il s'agit dans cette phase de calculer la métrique de la confiance, qui peut être modéliser par un graphe dont les noeuds sont les personnes et la pondération des arcs entre les noeuds représente le niveau de confiance, sachant que la confiance entre deux personnes n'est pas obligatoirement symétrique.
- **Le modèle de calcul de l'intérêt :** dans cette partie il s'agit de trouver la méthode pour en déduire la confiance entre les personnes. Divers algorithmes sont proposés dans ce sens dont la plupart se basent sur la propagation de la confiance ou bien l'agrégation de la confiance. En d'autres termes ces modèle assument que la confiance est transitive.

Parmi les systèmes à base de la confiance les plus répandus et qui ont eu beaucoup de succès sur internet nous citons : FilmTrust [63] pour la recommandation des films et TrustedOpinion [64] dédié à la recommandation des restaurants, cafés, bars et films. Les système de recommandation à base de confiance, exploitent les réseaux sociaux pour définir la notion de confiance grâce aux relations d'amitiés décrites dans le

paraphe social. Ainsi les recommandations sont générées automatiquement en fonction des recommandations des amis dans le réseau social et ce grâce à certaines valeurs comme l'appartenance aux mêmes communautés, les appréciations attribuées aux ressources, le suivi des personnes etc.). Par conséquent le développement de ces techniques à base de la confiance exige nécessairement la croissance des réseaux sociaux. La définition triviale de la confiance est l'amitié, ce qui revient à dire la confiance entre deux personne peut être définie comme le niveau de similarité de préférences entre les personnes. Ceci dit qu'un filtrage collaboratif classique peut aussi être considéré comme un système à base de confiance. Parmi les système qui se basent sur notion de similitude pour calculer la confiance nous citons SoNARS [65] et [66] où les auteurs analysent la corrélation entre le confiance et la similitude au niveau des intérêts dans une communauté en ligne. D'autres système comme [67] considèrent la notion de réputation comme un indicatif de confiance globale acquise par une personne sur la base de son comportement dans le passé et sans que ça ait un rapport avec les croyances des deux personnes pour lesquelles on calcule la confiance. une personne de bonne réputation peut être ainsi une personne qui a déjà donné des conseils pertinents.

De nombreux travaux ont tenté de prouver l'efficacité des méthodes de recommandation à base de la confiance à travers divers approches. Dans [68] les auteurs de ont choisi d'évaluer trois systèmes de recommandation de livres dont RatingZone, Sleeper et Amazon, et trois systèmes de recommandation de films à savoir Reel, MovieCritic, et Amazon pour enfin arriver à la conclusion que les utilisateurs ont une préférence pour les recommandations générées par les gens de confiance (les amis, la famille) plutôt que celles issues des systèmes de recommandation générale. Ainsi plusieurs chercheurs se sont focalisés sur les modèles de confiance comme un moyen pour supporter et compléter les approches classiques par filtrage collaboratif. les auteurs de [69] proposent architecture qui combinant trois types de recommandation : la confiance, filtrage collaboratif et de filtrage à base du contenu, ainsi grâce au agents le système proposent à un utilisateur des documents en profitant à la fois des avantages du filtrage collaboratif et celui à base du contenu et en incluant la confiance calculée à partir des notations des autres utilisateurs pour plus de précision. D'autres approche introduisent le filtrage à base d'opinion. Dans [70] les auteurs affirment que la confiance doit être issue à base de la similitude entre utilisateurs, ce qui implique que les amis sont exactement ces gens qui ressemble à notre vraie nature. Toutefois, le

modèle des auteurs est considéré comme une évolution du filtrage collaboratif en intégrant des agents intelligent. Cependant pour tenter d'améliorer les recommandations d'autres travaux utilisent des modèles qui introduisent la confiance entre les personnes par rapport à une ressource (la confiance au niveau d'objet ou item-level trust) [67]. Ainsi ces modèles améliorés prennent en considération le niveau de connaissance de chaque personne sur un domaine ou sur les ressources à recommander.

2.2 Le Web sémantique : structuration et valorisation des données

Le Web sémantique est venu avec de nouvelles pratiques concernant l'organisation des contenus Web, et une nouvelle infrastructure permettant aux agents logiciels d'aider efficacement les internautes dans leurs accès aux sources d'information et aux services [71]. Il s'agit d'arriver à un Web intelligent, où les informations ne seraient pas juste stockées mais comprises par les machines afin d'apporter des réponses pertinentes à l'utilisateur. Le Web sémantique avec ses outils a pour objectif de rendre automatique les requêtes, de réutiliser les données au travers d'applications diverses. En d'autres termes de libérer les données afin de fournir un moyen standardisé d'utiliser et de publier des données obtenues avec un minimum de travail. Le Web actuel étant essentiellement syntaxique, la structure des ressources est bien définie, mais le contenu reste très difficile à exploiter par les machines. Seuls les humains ont la capacité d'interpréter les contenus. Or la vraie valeur du Web est lié à ces données, ce qui fait que la représentation sémantique des contenus facilitera énormément la recherche de l'information ainsi que sa compréhension. Grâce à ses technologies le Web sémantique permettra aux utilisateurs de profiter du plein potentiel du Web, ainsi ils pourront chercher, trouver, partager et aussi combiner des informations plus facilement.

A l'heure actuel les internautes ont la capacité de manipuler et d'interagir avec divers média sociaux à savoir les forums, les réseaux sociaux, les blogs etc. Ils sont en mesure de faire des recherches ou même d'effectuer des achats en ligne en toute facilité. Pourtant, il serait préférable que la machine facilite le travail de l'homme et lui épargne l'effort en effectuant la majorité des tâches à sa place. Mais jusqu'à maintenant, les machines ont toujours besoin de l'assistance de l'humain, et les pages

Web dont leur conception sont plus faite pour être lisible et interprétable par des humains plutôt que des machines. Cependant l'objectif de base du Web sémantique est que les machines atteignent une certaine maturité et réalisent des tâches fastidieuses. Ainsi, le challenge du Web moderne ne se restreint pas à la contribution de ses utilisateurs humains, mais également à la contribution d'agents logiciels. Des agents intelligents qui proposent des agrégations et des techniques de représentions de données permettant de cacher d'une part la complexité des technologies pour l'utilisateur,et d'autres part d'apporter des réponses pertinentes aux questions et aux requêtes spécifiques. Notamment, expliciter les différents chemins et connexions entre les données favorise la découverte et la production de nouvelles informations et de nouvelles connaissances et des services à valeur ajoutée. Le Web sémantique vise une optimisation des contenus publiés sur le Web afin de faire coopérer l'homme et la machine efficacement. Et permettra aussi bien aux hommes qu'aux machines, ou agents logiciels, de comprendre le sens des données à travers des standards et des normalisations pour une meilleure utilisation. Nous passons donc d'un contenu orienté vers les hommes à un contenu constitué de données exploitables par les agents logiciels.

Dans la section suivante nous abordons les principes du Web sémantique ainsi que les différentes technologies liés aux différentes couches.

2.2.1 Web sémantique et Web de données

Selon Tim Berners-Lee [2], avant que le Web existe, Il a fallu ouvrir les différents programmes avec lesquels avaient été écrits les différents documents numérisés afin de les lire successivement. Le Web a simplifié le problème en inventant un langage unique HTML Hyper Text Markup Langage [16], qui permet de lire toutes sortes de documents et de les relier. Avec le Web sémantique, la philosophie est la même il crée un lien automatique pour relier les données stockées dans les différents fichiers et bases de données de nos ordinateurs. [2] synthétise le Web sémantique comme étant un Web où des ordinateurs utilisant des agents intelligents qui analysent toutes les données sur le Web (contenu, liens, transactions), qu'elles soient en langage naturel ou non. En d'autre termes « on passe d'un Web 2.0 qui exploite l'intelligence collective des hommes à un Web qui exploite l'intelligence collective des capteurs et des

16. http://www.w3.org/MarkUp/

données ».

Le Web sémantique entend alors remplacer le "Web des documents" par le "Web des données" d'où l'appellation Web de données qui veut également désigner le Web sémantique. Plus concrètement l'infrastructure du Web sémantique identifie et transforme les ressources de manière vigoureuse tout en renforçant l'aspect d'ouverture de données. Elle doit également assurer l'interopérabilité et faciliter la mise en oeuvre des calculs et des raisonnements complexes tout en maximisant leur validité. Et doit aussi offrir des mécanismes de protection à savoir les droits d'accès, d'utilisation et de reproduction, ainsi que des mécanismes permettant de qualifier les connaissances afin d'augmenter le niveau de confiance des utilisateurs. Mais la définition du Web sémantique ne se restreint pas à son infrastructure. Ce qui modélise ce concept c'est plutôt les applications développées sur cette infrastructure.

La manipulation des ressources Web par des machines requiert l'expression ou la description de celles-ci. A cet effet plusieurs langages sont donc définis, ils doivent permettre d'exprimer les données et les métadonnées, de décrire les services et leur fonctionnement et de disposer d'un modèle abstrait de ce qui est décrit grâce à l'expression d'ontologies. On présente ci-après les couches et principaux langages du Web sémantique figure 2.2. Nous reviendrons un peu plus loin sur les détails de chacune des couches et leurs apports dans cette nouvelle vision de l'architecture du Web.

Étant donné la quantité abondante des données disponibles sur le Web et leur caractère hétérogène et désordonné. Le but du Web sémantique est principalement une meilleure structuration des données ainsi qu'une affectation optimale des métadonnées afin de simplifier leurs interprétations par les machines. Pour mener à bien ces objectifs, le W3C a mis en place, progressivement, une série de recommandations pour l'ensemble des couches de l'architecture. La figure 2.3 récapitule les standardisations faite par le W3C au sujet du Web sémantique, au cours du temps.

2.2.1.1 La couche URI et unicode

L'URI (Uniform Resource Identifier) représente un aspect central de l'infrastructure, d'où sa raison d'être à la base des couches l'architecture du Web sémantique. C'est un protocole simple et extensible permettant d'identifier de manière unique et uniforme toute ressource sur le Web, c'est aussi un moyen pour que les données soient identifiables par la machine. L'URI dérive des concepts introduits par l'initiative de globalisation de l'information dans le World Wide Web. L'URI est une séquence de

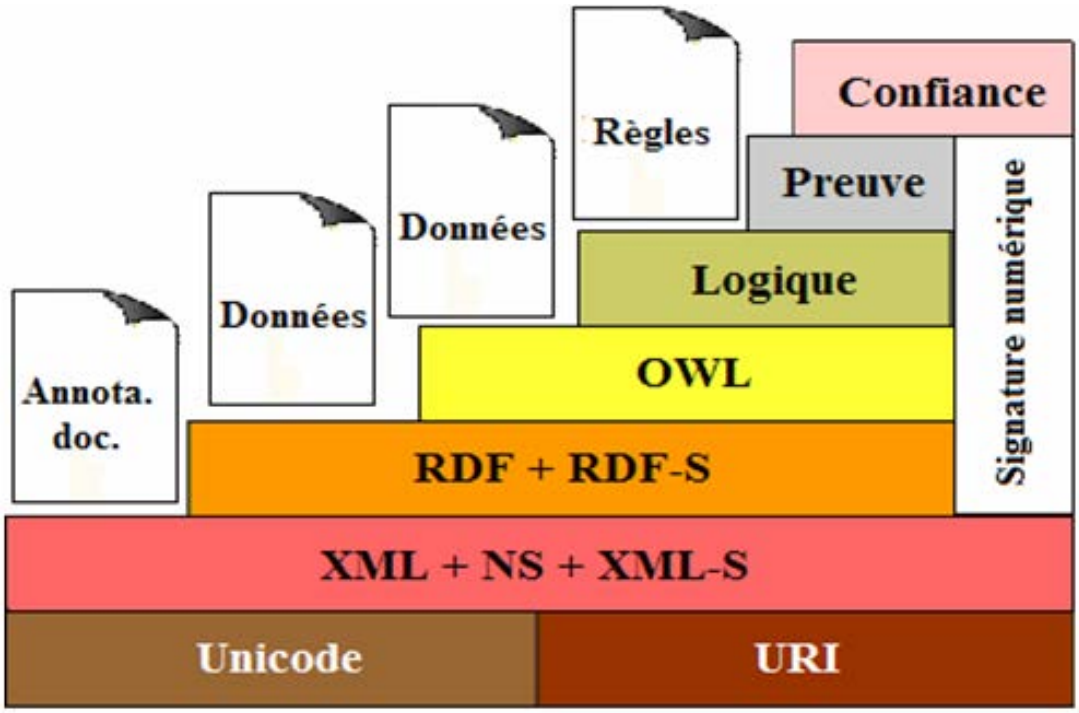

FIGURE *2.2: Architecture en couches du Web sémantique [2]*

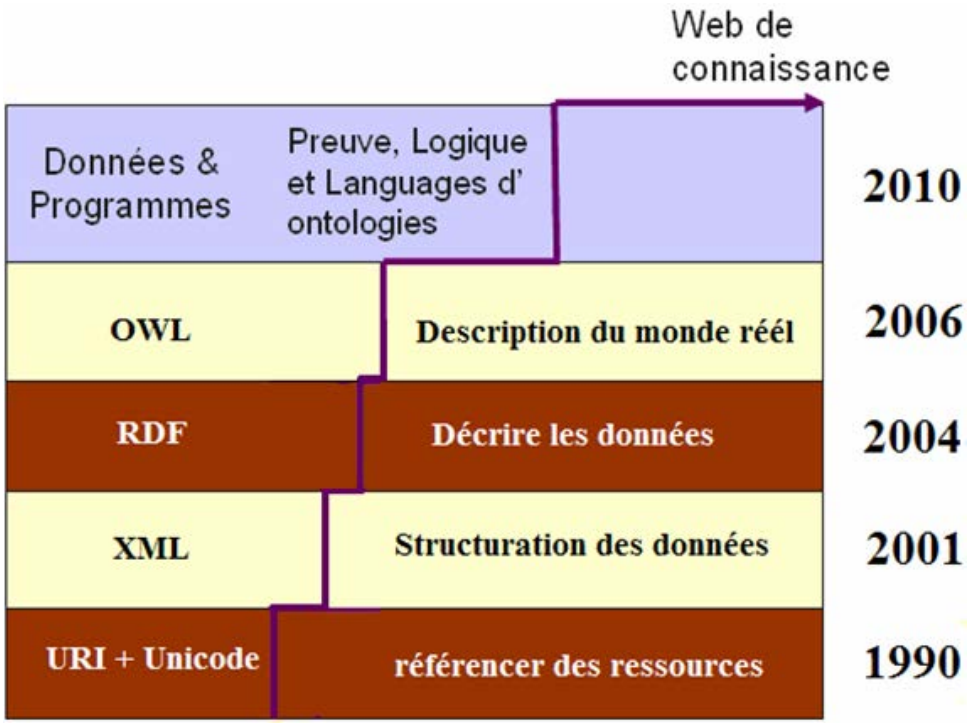

FIGURE *2.3: Évolution dans le temps des différentes couches du web sémantique. [2]*

caractères avec une syntaxe restreinte, qui permet d'identifier toute ressource utilisée dans le cadre d'une application Web sémantique. La notion d'uniformité liée à l'URI permet d'abord à différents types d'identificateurs de ressources d'être utilisés dans le même contexte, même si le mécanisme qui leur permet d'y accéder est différent. Ensuite elle assure une interprétation sémantique uniforme des conventions syntaxiques communes pour les différents identificateurs. Cette uniformité permet également la réutilisation de ces identificateurs dans différents contextes.

Par ailleurs, il est à noter que les données sont toujours encodées avec le jeu de caractères Unicode pour un maximum d'interopérabilité. C'est pourquoi cet élément figure dans cette couche de bas niveau, au même titre que l'URI.

2.2.1.2 La couche XML, NS, XML Schema

A ce niveau de l'architecture, Il n y a toujours pas la notion d'affectation de la sémantique à l'information. Il s'agit seulement d'une couche syntaxique, de bas niveaux, qui permet de structurer les données et organiser selon un format de message standard, et ce grâce au langage de balisage extensible XML (eXtensible Markup Language). En d'autres termes, XML permet d'indiquer l'organisation logique du contenu d'un document, mais n'assure la sémantisation de l'information [17]. A cette phase la structuration de l'information, revient principalement à séparer la mise en forme et le contenu.

2.2.1.3 La couche RDF et RDF Schema

Après le référencement des ressources avec le protocole URI et la structuration des informations avec le langage XML, l'étape suivante consiste à les annoter, afin de les doter d'un sens interprétable par la machine. C'est justement le rôle de la couche RDF et RDF-S dans l'architecture du Web sémantique.

RDF (Resource Definition Framework) est un langage permettant de décrire des ressources web grâce à des triplets de la forme : (Sujet, prédicat, objet).

- **Sujet :** la ressource à laquelle s'applique le prédicat.
- **Prédicat :** une propriété qui s'applique au sujet.
- **Objet :** une ressource liée au sujet par le prédicat.

17. http://lespetitescases.net/structurer-decrire-et-organiser-l-information-1

C'est un modèle standard pour l'échange de données sur le "Web des données". Il facilite la fusion des données même quand les schémas sous-jacents diffèrent.

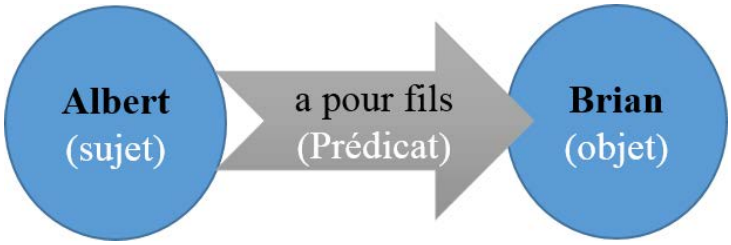

FIGURE *2.4: Exemple de triplet RDF*

Le prédicat doit nécessairement être identifié par une URI, ce qui n'est pas forcément le cas pour le sujet et l'objet qui peuvent être aussi des noeuds anonymes. Nous avons à disposition de nombreuses syntaxes pour écrire des documents RDF dont la plus utilisée est RDF/XML qui est une sérialisation de RDF en XML.

RDF Schéma RDF Schéma est une extension de RDF. C'est un méta-vocabulaire, c'est-à-dire un vocabulaire qui permet d'en décrire d'autres. RDFS 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. Pour cela, RDFS ajoute les notions de « domain », correspondant au domaine de définition d'une fonction, et le « range », son ensemble d'arrivée. RDFS fournit des éléments de bases pour la définition d'ontologies ou vocabulaires destinés à structurer des ressources RDF. Les principales caractéristiques de RDFS :

- **rdfs :Class** permet de déclarer une ressource RDF comme une classe pour d'autres ressources.
- **rdfs :subClassOf** permet de définir des hiérarchies de classes.
- **rdfs :domain** définit la classe des sujets liée à une propriété.
- **rdfs :range** définit la classe ou le type de données des valeurs de la propriété.

2.2.1.4 La couche ontologique

Une ontologie désigne à la fois une modélisation et la structure qu'elle modélise. Le terme ontologie est emprunté de la philosophie où il fait référence à la science qui « étudie l'être en tant qu'être ». Prenant l'exemple d'un texte, on s'intéresse autant à la structure qu'au sens. Cette couche utilise le langage OWL est une extension de

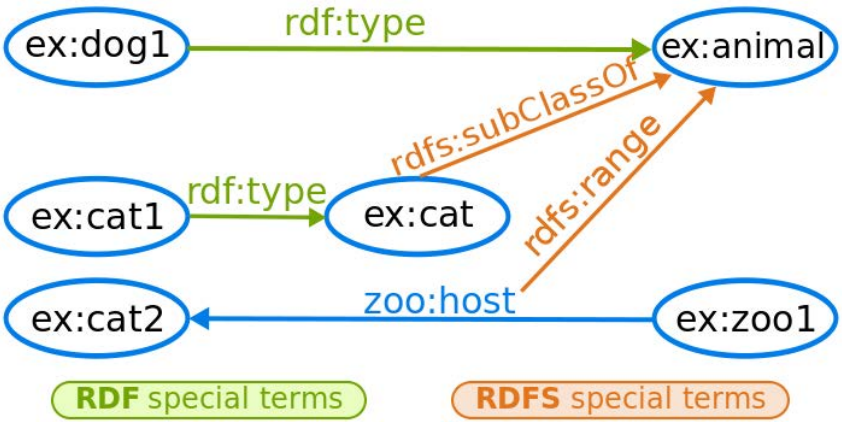

FIGURE *2.5: Exemple de graphe RDF*

RDF Schémas qui est basé sur RDF. Il définit un vocabulaire riche pour décrire les ontologies. Les ontologies sont définies comme :

Définition 3 *« Une spécification formelle et explicite d'une conceptualisation partagée » [72] plus précisément.*

- ***spécification formelle :*** *compréhensible par une machine*
- ***spécification explicite :*** *les concepts, relations, fonctions, contraintes, axiomes sont explicitement définis*
- ***conceptualisation :*** *modèle abstrait d'une partie du monde que l'on veut représenter*
- ***partagée :*** *les connaissances représentées sont partagées par une communauté*

L'ontologie est une modélisation du monde réel en concepts et relations entre ces concepts. Dont les principales composantes sont les suivantes [73] :

- **Concept :** C'est la représentation abstraite des éléments du domaine. On peut également les appeler termes ou classes. Ces concepts peuvent être classés selon différents critères dans la taxonomie (niveau d'abstraction, atomicité, etc.).
- **Relations :** Elles expriment les associations entre les différents concepts définis dans la taxonomie. Les différents types de relations qui peuvent exister sont : « Spécialisation/Généralisation », « Agrégation ou Composition », « associé à », « composé de », etc.
- **Fonctions :** Il s'agit des relations particulières où un élément est défini par les n-1 autres éléments.

– **Axiomes :** Constituent des assertions considérées toujours comme vraies.
– **Instances :** Ce sont des exemples particuliers de concepts.

La figure 2.6 représente un exemple d'une ontologie élémentaire.

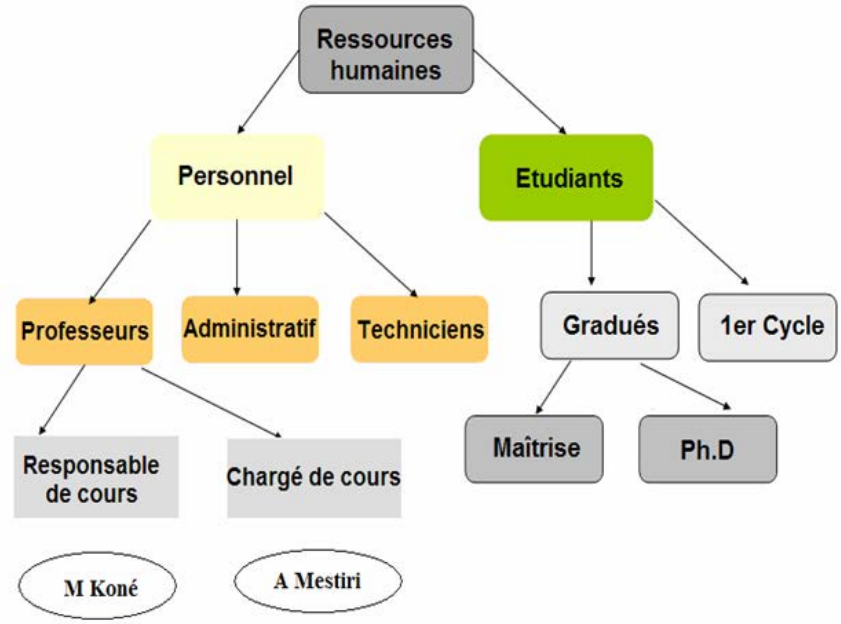

FIGURE *2.6: Exemple d'une ontologie élémentaire du domaine.*

La construction d'une ontologie commence par l'analyse des besoins, jusqu'à à la validation et l'évaluation. La Figure 2.7 une vue exhaustive des différentes phases de la génération de l'ontologie.

Comme les ontologies sont utilisées comme des composantes de systèmes logiciels, alors leur développement s'appuie sur les mêmes principes du génie logiciel [74]. Le cycle de vie d'une ontologie passe alors par les étapes suivants : spécification, conceptualisation, formalisation, intégration, implantation, et maintenance. Détailler la couche ontologique, revient implicitement à évoquer les langages ontologiques, dont le plus connu est le OWL (Ontology Web Language).

OWL : Ontology Web Language

Définition 4 *OWL est une extension de RDF Schémas qui est basé sur RDF. Il définit un vocabulaire riche pour décrire les ontologies. Ce langage est recommandé par le W3C comme un standard pour le web sémantique depuis 2004 et constitue un pilier*

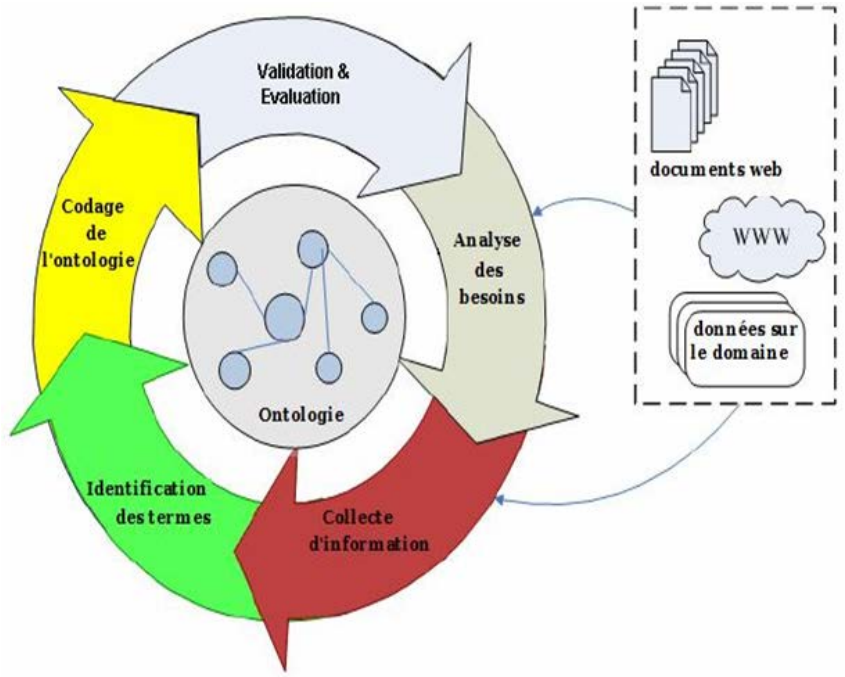

FIGURE *2.7: Les différentes étapes de génération de l'ontologie*

pour le web sémantique, selon Tim Berners-Lee [2]. OWL a pour principale caractéristique de décrire les concepts et les relations entre ces concepts, pour permettre un raisonnement logique au sein des systèmes d'information, appelé inférences.

Le langage OWL peut être défini en trois sous langages(La figure 2.8), suivant le niveau d'expressivité qu'on cherche à exprimer.

- **OWL Lite :** Il s'agit d'un sous-ensemble de OWL qui permet d'exprimer la classification, et les relations simples entre les classes. Ce sous langagene permet pas d'exprimer des contraintes complexes sur les classes ou sur les associations.
- **OWL DL :** Permet un meilleur niveau d'expressivité tout en maintenant la complétude et la décidabilité (tous les calculs doivent être achevés en un temps fini). Ce sous-ensemble repose sur les caractéristiques de la logique de description pour inclure des propriétés utiles aux systèmes de raisonnement.
- **OWL Full :** Ce sous-ensemble offre un maximum d'expressivité, mais sans aucune garantie de calcul. OWL permet donc à une ontologie d'augmenter le sens du vocabulaire prédéfini.

FIGURE *2.8: Les sous langages OWL*

OWL 2 : Ontology Web Language 2 [75] [3]

Définition 5 *[3] OWL 2 est un langage ontologique pour le Web Sémantique possédant une signification formellement définie. OWL 2 est une extension et une révision du langage d'ontologies Web OWL. OWL 2 est conçu pour faciliter le développement d'ontologies et leur partage via le Web, avec pour objectif final la plus grande accessibilité des contenus du Web pour les machines. Les ontologies OWL 2 intègrent des classes, des propriétés, des individus et des valeurs de données et sont stockées dans des documents Web Sémantique.*

La Figure 2.9 donne une vue d'ensemble du langage OWL 2 en montrant ses principaux composants et la manière dont ils sont associés. L'ellipse au centre représente la notion abstraite d'une ontologie qui peut être vue soit comme une structure abstraite, soit comme un graphe RDF . Dans la partie supérieure sont représentées des syntaxes concrètes qui peuvent être utilisées pour sérialiser et échanger des ontologies. Dans la partie inférieure sont représentées deux spécifications sémantiques définissant la signification des ontologies OWL.

Les éléments de bases de la structure OWL 2 [3]

- **les ontologies :** les ontologies OWL 2 peuvent être utilisées de concert avec des informations décrites à l'aide de RDF et les ontologies OWL 2 sont elles-mêmes principalement échangées sous la forme de documents RDF.
- **la syntaxe :** une syntaxe concrète est nécessaire pour stocker des ontologies OWL 2 et les échanger entre les outils et les applications. La syntaxe d'échange principale pour OWL 2 est RDF/XML. D'autres syntaxes concrètes peuvent également être utilisées. Parmi ces sérialisations se trouve :

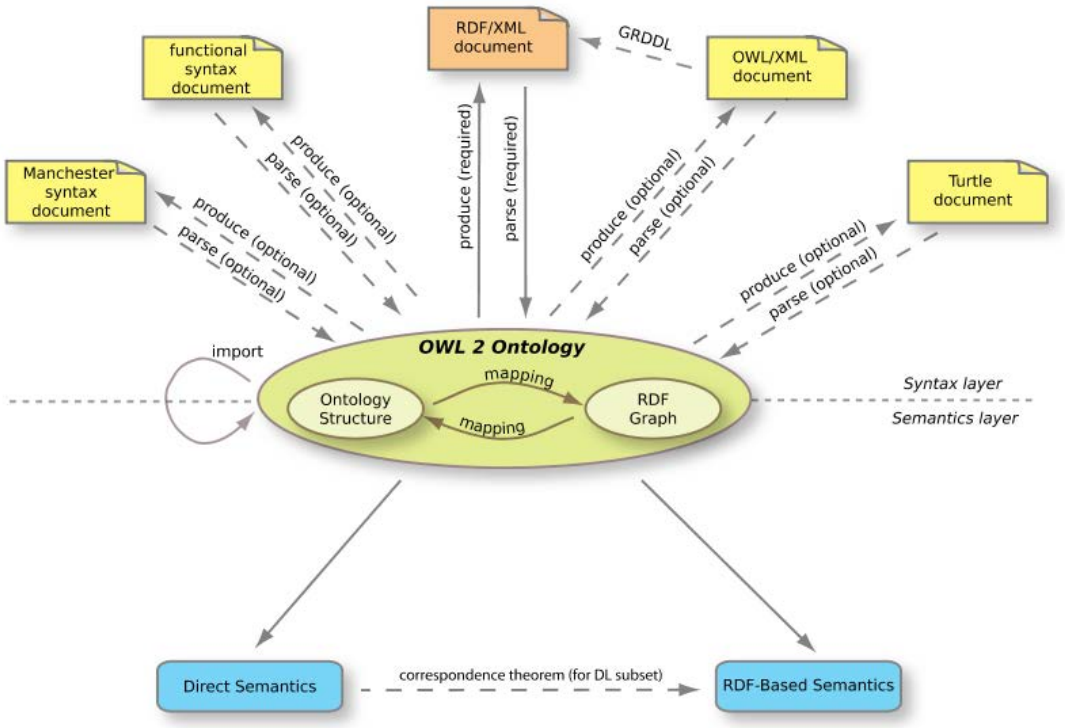

FIGURE *2.9: La structure de OWL 2 [3]*

- Turtle : permettant de lire/écrire plus facilement des triplets RDF.
- OWL/XML : plus simple à traiter en utilisant des outils XML.
- Syntaxe fonctionelle : permettant de mieux voir la structure formelle des ontologies.
- Syntaxe Manchester : permettant de lire/écrire plus facilement des ontologies DL.

– **la sémantique :** Le document de spécification structurelle de OWL 2 définit la structure abstraite des ontologies OWL 2 mais il ne définit pas leur signification. Deux alternatives pour associer une signification à des ontologies OWL 2
  - **La sémantique directe OWL 2 :** cette sémantique assigne une signification directement dans les structures ontologiques. Elle est compatible avec le modèle sémantique théorique de la logique de description SROIQ et ainsi elle ne peut être appliquée qu'a des ontologies satisfaisant certaines conditions syntaxiques, ces dernières sont appelées ontologies "OWL 2 DL".
  - **La sémantique OWL 2 basée sur RDF :** cette sémantique tend les conditions sémantiques définies pour RDF. Elle attribue une signification directement sur

les graphes RDF et donc indirectement sur la structure ontologique via la correspondance vers les graphes RDF. La sémantique basée sur RDF peut être appliquée à toute ontologie OWL 2, sans restrictions, comme toute ontologie OWL 2 peut avoir une correspondance vers RDF.

- **les profils :** Les profils OWL 2 sont des sous-langages ou sous-ensembles syntaxiques de OWL 2 qui offrent d'importants avantages dans certains scénarios d'applications. ils sont définis comme une restriction syntaxique de la spécification structurelle de OWL 2. Trois profils différents sont définis :
  - **OWL 2 EL :** ce profil permet d'obtenir des temps d'exécution polynomiaux pour toutes les tâches de raisonnements standards ; ce profil est particulièrement approprié lorsque des ontologies très volumineuses sont nécessaires, et lorsqu'il est possible de faire des concessions sur l'expressivité en échange de garanties sur les performances.
  - **OWL 2 QL :** ce profil permet aux requêtes conjonctives d'être traitées avec une complexité logarithmique en utilisant les technologies des bases de données relationnelles standards. Il est particulièrement approprié pour des applications où des ontologies relativement légères sont utilisées pour organiser un grand nombre d'individus et où il est utile ou nécessaire d'accéder aux données directement via des requêtes relationnelles.
  - **OWL 2 RL :** ce profil permet l'implémentation d'algorithmes de raisonnement en temps polynomial utilisant les technologies d'extension de règles des bases de données opérant directement sur les triplets RDF. Il est particulièrement approprié pour des applications où des ontologies relativement légères sont utilisées pour organiser un grand nombre d'individus et où il est utile ou nécessaire d'opérer directement sur les données sous la forme de triplets RDF.

Pour résumer la structure OWL 2 est très similaire à celle de OWL. Les nouvelles fonctionnalités OWL 2 concernent principalement des simplifications syntaxiques (la classe union de classes disjointes), et l'augmentation de l'expressivité (les clés, les chaînes de propriétés, des types de données plus riches, plage de données, les restrictions de cardinalités qualifiées, les propriétés disjointes asymétriques et réflexives, et des possibilités d'annotations améliorées). OWL 2 a également défini trois nouveaux profils (OWL 2 EL, OWL 2 QL, et OWL 2 RL) et une nouvelle syntaxe (Syntaxe Manchester).

2.2.1.5 La couche logique

La couche logique est utilisé pour exprimer les règles d'inférences. Cette couche repose sur les langages ontologiques dans l'architecture. La logique est la discipline qui étudie les principes et les formes du raisonnement, la logique de description est celle qui est, généralement, la plus adoptée pour la représentation des règles d'inférences.

Définition 6 *La logique de description est définie comme étant une famille de formalismes de représentation de la connaissance basée sur la logique. Elle est conçue pour représenter et raisonner sur la connaissance d'un domaine d'application d'une manière structurée et bien comprise. Elle dérive des réseaux sémantiques. [76]*

Les deux concepts de base dans les logiques descriptives sont : les concepts (prédicats) et les rôles (relations binaires), et les notions les plus importantes à prendre en considération sont la **satisfiabilité** et la **subsomption**. Un concept est dit satisfaisable quand son expression ne réfère pas à un ensemble vide. En d'autres termes s'il existe une interprétation possible pour ce concept. La Satisfiabilité est un cas particulier de la subsomption, qui n'est autre que la relation de généralisation/spécialisation. C'est pourquoi on parle d'une relation entre la satisfiabilité et la subsomption).

Les deux éléments principaux dans une architecture d'inférence basée sur la logique de description sont le TBox (T axonomy Box) qui représente la connaissance conceptuelle du domaine et englobe un ensemble de formules relatives aux informations terminologiques. Le ABox (Assertional Box) quant à lui, définit les instances, c'est-à-dire l'ensemble de formules relatives aux informations sur les assertions. Celles-ci sont instables et dépendantes du domaine. Ainsi, toutes les informations connues sont alors modélisées comme un couple <T,A> [77]. La figure 2.10 représente l'architecture d'une inférence basée sur la logique de description.

2.2.1.6 La couche confiance et preuve

La prochaine étape dans l'architecture est la couche **preuve**, très peu de choses sont dites sur cette couche de l'architecture car contrairement aux trois premières couches déjà été standardisées et recommandées par le W3C. Cependant très peu d'auteurs évoquent dans leurs littératures des détails sur couches de haut niveau. En effet, les recherches menées jusqu'à présent, dans ce contexte,sont plutôt focalisées

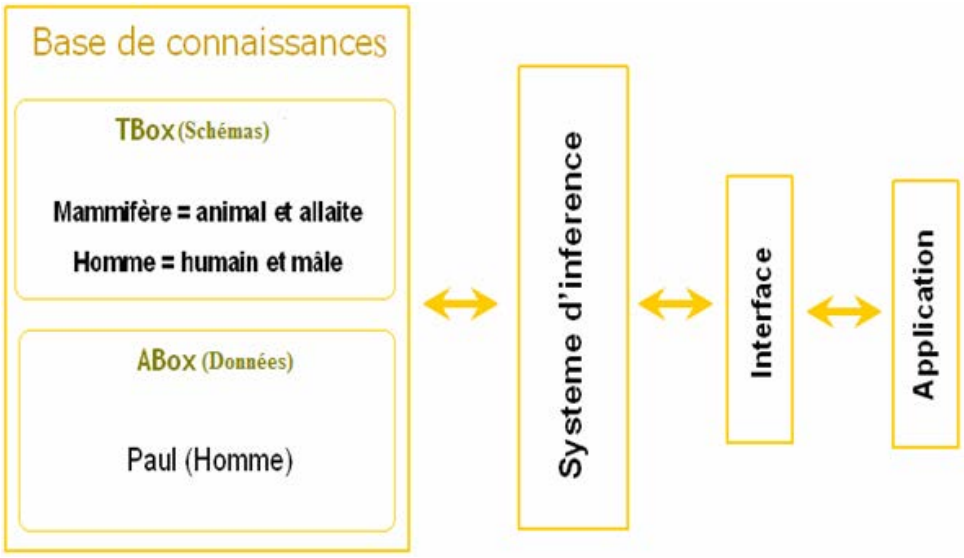

FIGURE *2.10: Architecture d'inférence basée sur la logique de description*

sur le référencement des ressources, la structuration de l'information (XML), l'affectation des métadonnées (RDF), la modélisation du monde réel sous forme d'ontologies (OWL), en faisant abstraction de la sécurité et des preuves qui servent fondamentalement à justifier la pertinence de l'information.

La couche preuve a pour objectif de prouver la pertinence de l'information retournée par les couches de plus bas niveau et des déductions obtenues à partir des inférences. Selon les recherches menées il n'existe encore aucun langage de preuves standardisé par le W3C. De ce fait un langage de preuve constitue un moyen simple pour prouver si une déclaration est juste ou pas. Une instance de ce dernier consiste en général en une liste de toutes les étapes d'inférence par lesquelles a transité l'information en question.

Le Web est un environnement ouvert très dynamique, où chacun est en mesure d'éditer et de publier des informations de façon très simple. La couche Confiance a pour fin l'évaluation de la fiabilité de l'information et des raisonnements. Cette couche repose sur les signatures numériques, le cryptage des données et sur la fiabilité des sources d'information (agents de confiances, certifications, etc.) [78].

2.2.2 Le Web sémantique pour l'analyse des données sociales et la prédiction des liens

Au-delà de la déclaration d'amitié et de la mise en ligne de contenus et d'informations personnels, les services de réseautage deviennent de plus en plus une place d'échange et de création de l'information. De nombreux travaux s'intéressent à l'analyse des données sociales pour la recommandation des contenus. Les média sociaux génèrent une grande quantité de données sous forme de statuts, de commentaires, hashtags, etc. L'enjeu est de tirer profit de cette masse d'information pour faire des recommandations. Plus précisément il s'agit d'apprendre des connaissances à partir de traces laissées par les utilisateurs des réseaux sociaux, pour les exploiter au sein d'un environnement d'apprentissage. Les nouveaux challenges ne se restreignent pas à la contribution des utilisateurs humains, mais également à la contribution d'agents logiciels et de nouvelles technologies sémantiques qui parcourent et valorisent les données. Le Web sémantique vise l'exploitation systématique de la sémantique des ressources, et la mise en évidence des différentes connexions et chemins disponibles entre des données, favorisant la découverte et la production de nouvelles informations et de nouvelles connaissances [79]. Le Web Social peut représenter un bon support pour la création de données formalisées selon les principes du Web sémantique. Les technologies sémantiques peuvent assurer la structuration et l'homogénéité des données produites par les différents médias sociaux.

Les processus d'apprentissage doivent être efficaces et rapides. Ceci nécessite un contenu adéquat aux besoins spécifiques et par conséquent des mécanismes puissants pour l'organisation et le tri des ressources pédagogiques. Les perspectives envisagées par l'inclusion du Web sémantique en éducation sont de nouvelles méthodes et de nouveaux instruments pour améliorer les capacités des plates-formes existantes. Pour qu'elles s'adaptent au rythme de leurs usagers et contribuent efficacement à l'évolution de leurs connaissances et de leurs compétences [80]. De plus le Web Sémantique permet d'identifier les compétences qu'un cours particulier vise à développer chez les apprenants. Autrement dit l'apprenant ne doit pas se soucier de chercher les bons cours adaptés à son niveau et permettant d'améliorer ses compétences, le système doit être en mesure de le faire pour lui. Autant faire des synthèses et des conclusions pertinentes sur le niveau de performance des apprenants, les difficultés qu'ils rencontrent afin d'améliorer les contenus qu'on leur recommande.

Notre approche a pour but de recenser des informations pertinentes sur les appre-

nants via leurs contributions, et principalement via leur activité de social tagging. En effet aujourd'hui le moyen le plus utilisé par les internautes pour la classification des ressources partagés sur les différents services en lignes. Nous avons choisi d'exploiter les hashtags, des écrits générées par les utilisateurs d'une pratique collaborative spontanée et ne demandant aucune compétence particulière. Cette activité de tagging engendre énormément de données mais elles restent ambiguës et difficilement exploitables en l'absence d'une vraie analyse sémantique. Pour bien tirer profit de ce phénomène social, les technologies sémantiques sont le moyen idéal.

Nous concluons donc que les réseaux sociaux sont des moyens privilégiés d'informations abondantes et souvent intéressantes, dans le sens où elles décrivent les tendances, orientations, préférences et centres d'intérêts de leurs producteurs. Le Web sémantique dispose des moyens pour formaliser, réutiliser et mettre en valeur les informations et les liens des médias sociaux. Le Web social et le Web sémantique ne sont pas concurrents mais plutôt complémentaires. Une cohabitation des deux paradigmes du Web est le meilleur moyen d'augmenter les capacités d'accès à une information pertinente et personnalisée. Notre approche vise à profiter des écrits et données partagées sur les réseaux sociaux pour l'élaboration de folksonomies hiérarchisées et enrichies sémantiquement. Ces derniers nous permettront d'enrichir le profil de l'apprenant avec un ensemble d'intérêts générés à partir des hashtags sémantisés et ainsi de proposer des ressources pédagogiques en adéquation avec les intérêts de l'apprenant. Dans la section qui suit, nous évoquons l'enrichissement sémantique des hashtags et les différents travaux autour de cette question.

2.2.3 L'enrichissement sémantique des hashtags

Plusieurs travaux ont tenté d'exploiter sémantiquement les données sociales pour générer des modèles structurés et enrichis sémantiquement. Une première catégorie de travaux visent à extraire la sémantique émergente des folksonomies en mesurant la similarité sémantique des hashtags[81] ou en exploitant les associations de hashtags via les utilisateurs pour extraire des relations taxonomiques[82]. D'autres approches s'appuient davantage sur les contributions des utilisateurs pour taguer les hashtags [83], ou pour structurer les hashtags à l'aide d'une syntaxe simple permettant de spécifier des relations de subsomption ou de synonymie[84]. Plus récemment des solutions visant à intégrer le plus naturellement possible les interfaces de tagging, et les systèmes à base d'ontologies ont émergé. A savoir les ontologies SCOT pour les ha-

shtags ou SIOC pour les sites sociauxqui permettent d'améliorer l'interopérabilité des plateformes d'échanges et production de connaissances. Nous notons également l'ontologie du tagging "NiceTag" [85] où les auteurs proposent un modèle pivot permettant d'intégrer divers modèles de hashtags tout en prenant en considération la diversité de forme et d'usages des hashtags. De leur côté les auteurs de [86] proposent MOAT, un modèle permettant de relier les hahtags avec des URIs de resources du Web décrivant leur signification.

D'un autre coté nous trouvons des approches qui intègrent les mesures de similarité dans un processus de mapping entre les hashtags et les concepts d'ontologies disponibles en ligne [87]. Et [88] pose qu'il n'y a pas d'opposition entre les ontologies et les folksonomies et propose de construire une "ontologie de folksonomie" qu'il a nommé la "TagOntology" un projet de construction d'une ontologie commune dédié à la formalisation et la conceptualisation de l'acte du tagging, ce modèle met en oeuvre quatre entités pour caractériser un évènement de "tagging" : la ressource ; le tag ; l'utilisateur ; et le domaine au sein duquel le tagging s'inscrit. [88] proposait également de "tagger les hashtags". Par ce biais, il serait possible d'indiquer que tel hashtag est synonyme de tel autre hashtag, ou encore que tel hashtag est adéquat ou non pour tel objet. Les auteurs de [89] ont de leur côté appliqué l'idée d'exprimer les hashtags directement à l'aide des langages du Web sémantique. Les hashtags sont donc collectés autant que de nouvelles classes d'ontologie et servent à annoter les pages d'un wiki 'sémantique". Dans [90] les auteurs utilisent une folksonomie et des calculs de co-occurrences comme point d'entrée à la construction d'une ontologie via l'exploitation des graphes pour la récupération des liens sémantique entre concepts.

Plus récemment d'autres approches de structuration ont également émergé. [91] font l'adaptation d'une méthode dédiée à l'enrichissement d'ontologies et basée sur la recherche de motifs séquentiels pour la découverte de nouveaux liens labellisés, pour des fins de structuration de folksonomies. Via leur approche les auteurs visent le rapprochement entre tags, découvrir d'autres non existants dans la folksonomie, expliciter les relations entre tags, et typer ces relations en identifiant des relations d'hyperonymie, d'hyponymie, de synonymie ou bien en proposant des étiquettes pour décrire ces relations. Cette approche présente l'avantage d'être indépendante de la structure initiale de la folksonomie et des textes utilisés en entrée du processus. Autrement dans [92] ils adoptent une approche de structuration de la folksonomie qui consiste à spécifier des relations sémantiques de thesaurus entre les tags en suivant le standard

SKOS. Ils proposent un système qui aide les utilisateurs, via des suggestions automatiques, à maintenir leur propre structuration de la folksonomie tout en bénéficiant des contributions des autres utilisateurs.
Notre vision consiste à exploiter la dynamique participative, la réactivité et l'engagement des utilisateurs sur les structures sociales, pour faire une organisation des hashtags en ligne. Nous recherchons à travers cela de lever l'ambigüité de sens des hashtags, assigner aux hashtags des catégories ou thématiques, trouver les différents thèmes abordés dans une conversation donnée, mettre en avant les hashtags les plus importants qui nous donnerons une vision claire sur les tendances d'un internaute, et ce en se basant sur la meilleur mesure de similarité adapté à notre contexte. Par ailleurs l'enrichissement sémantique des hashtags nous permettra d'un côté de construire un riche dictionnaire de hashtags, contenant le maximum de hashtag avec leurs définitions, et d'un autre côté avoir une base de données riche sur laquelle nous pouvons faire des traitements algorithmiques et en déduire la classification des hashtags d'un utilisateur selon leur degré de correspondance. Le résultat sera donc un vocabulaire structuré et assurant une meilleur visibilité des utilisateurs et ainsi permettant l'enrichissement automatique de leurs profils.

2.3 Synthèse sur l'impact et les aspects de cohabitation du Web social et Web sémantique pour le e-learning

Dans cette section nous exposons les portées, les intérêts et l'influence des deux paradigmes Web social et Web sémantique sur les systèmes d'apprentissage en e-learning. Nous préconisons de soutenir la recommandation personnalisée dans le e-learning via l'analyse sémantique des hashtags contenus dans les données partagées par les apprenants sur les réseaux sociaux.

2.3.1 Du e-learning au e-learning 2.0

Le développement des technologies Web joue un rôle de plus en plus central dans l'évolution de diverses disciplines et affectent significativement les modes de vie. Sur le plan éducatif les changements se sont manifestés dans un premier temps par le

passage au e-Learning. Le e-Learning a été introduit comme une solution complémentaire à l'enseignement traditionnel en vue d'améliorer la qualité d'apprentissage. De nombreux systèmes éducatifs ont intégré le e-Learning dans leur cursus de formation, en commençant par le blended Learning [93] (université Lyon2 [18] , université de liège [19] ...), et en arrivant jusqu'à l'apprentissage entièrement à distance (université virtuelle canadienne [20] , Université Virtuelle Africaine [21], Campus virtuel palestinien [22] ...).

Nous sommes alors passés de l'apprentissage transmissible, où l'enseignant est le seul détenteur de l'information et le rôle de l'apprenant se limite à la réception des informations, vers un autre apprentissage centré apprenant et qui se détache de plus en plus des contraintes spatiotemporelles de l'enseignement classique. L'enseignement traditionnel ne donnait pas assez de temps à l'apprenant pour analyser les informations qu'il reçoit. Le e-learning a donné l'apprenant plus de liberté pour établir le lien entre les informations recueillis et ses prérequis, et ainsi construire sa propre vision en utilisant ses propres schémas mentaux. Le e-learning permet de faire disparaitre la passivité de l'apprenant grâce aux possibilités d'interactivité et de collaboration qu'il offre. Il a également étendu les possibilités de l'apprenant avec un panel de ressources considérablement élargie, ainsi qu'une flexibilité et adaptabilité accrues.

Cependant face à l'émergence du Web 2.0 et réseaux sociaux, les choses ont évolué, et de nouvelles caractéristiques des apprenants ont vu le jour. La réussite scolaire, qui se mesurait principalement en termes cognitifs, est déterminée dorénavant par les aptitudes techno-cognitives, c'est-à-dire la compétence des apprenants à maîtriser les technologies qui les entourent et à les mettre au service de leurs apprentissages plutôt que de les subir ou d'y réagir [94]. Nous sommes face à de nouveaux défis en ce qui concerne l'apprentissage en ligne avec une nouvelle génération native du numérique. La génération Y "*désignant les personnes nées entre 1978 et 1994 des jeunes et adolescents qui ont grandi au moment où l'Internet s'est généralisé*" [95].

Les apprenants d'aujourd'hui ont développé de nouvelles pratiques : générer du contenu, créer leurs propres objectifs et intérêts, contrôler leur propre contenu, partager et collaborer dans un intérêt commun... A la fois natifs du numérique et utilisant les outils du Web 2.0, on les nomme des "apprenants 2.0". Le e-learning a tracé le

18. http://www.univ-lyon2.fr
19. https://www.ulg.ac.be/cms/c_5000/fr/accueil
20. http://www.cvu-uvc.ca/englishFR.html
21. http://uva.ulb.ac.be
22. http://fr.unesco.org

chemin vers l'éducation personnalisée, mais nous sommes face à une génération demandeuse d'un apprentissage de très haute qualité, et qui cherche à personnaliser son expérience Web et se construire son propre environnement d'apprentissage.

L'un des problèmes du e-learning est bien souvent l'arrêt en cours de formation. Les apprenants ne vont pas au bout de leur apprentissage. En effet, seuls face à leurs écrans, ils ne trouvent pas la motivation pour achever ce qu'ils ont commencé. Le e-learning doit être mesure de s'améliorer pour lier les gens les uns aux autres afin qu'ils se parlent et s'auto-forment. Dans son expérience e-learning, l'apprenant 2.0 ne se contente pas de chercher l'information il cherche les communautés de pratiques. Subséquemment pour un e-learning actif, il va falloir prendre en considération la masse importante d'informations échangées sur les structures sociales. L'apprentissage doit intégrer une nouvelle dimension qui n'est rien d'autre que le social. Dans cette perspective le e-learning prend une nouvelle posture et ainsi une nouvelle appellation : "e-Learning 2.0". On peut définir e-learning 2.0 "*comme une stratégie e-learning qui utilise les technologies du Web 2.0, et dont la principale caractéristique est que les apprenants peuvent contrôler activement leur contenu et l'orientation d'apprentissage. Il inclut la communication, l'apprentissage collaboratif, les réseaux sociaux, et ainsi de nouveaux rôles pour les apprenants et les enseignants*" [96].

L'idée est de profiter des avantages de la communauté en pleine croissance sur les réseaux sociaux, au sein des environnements d'apprentissage. La figure 2.12 illustre l'idée d'établir des ponts de communication entre les contributions des apprenants sur les réseaux sociaux (l'informel) et celle sur les plates-formes d'apprentissage (le formel).

L'idée d'apprentissage social ne date pas d'aujourd'hui. Albert Bandura, psychologue canadien né en 1925, a déjà publié son ouvrage intitulé "Social Learning and Personality" en 1963 [97]. Il ne sera traduit en français que 25 ans plus tard. Bandura avait déjà conclu dans son ouvrage que tout apprentissage est social. Maintenant, l'avènement des réseaux sociaux a favorisé l'apprentissage collaboratif, et les échanges qui en découlent s'avèrent très riches et prometteurs, car les va et vient entre les membres et au sein d'une communauté permettent des expériences enrichissantes. On a vu l'émergence d'une nouvelle gamme d'outils qui allie le social au e-learning et qui figure principalement dans des réseaux sociaux spécialement axés sur le e-Learning. ELGG [98], Edublogs Campus [99], Audacity [100], MindMeister [101], Ning [102] sont des plateformes sociales promouvant l'apprentissage. Elles sont éga-

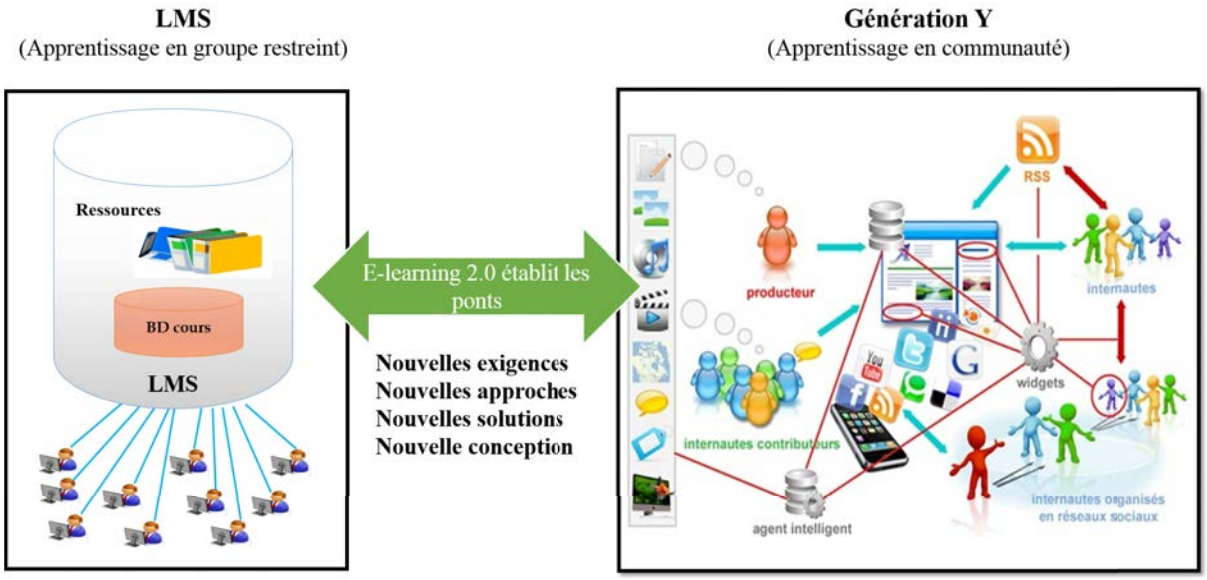

FIGURE *2.11: La generation Y change la posture de l'apprentissage*

lement open source, entièrement personnalisables, configurables et extensibles.

Comme cité précédemment les apprenants 2.0 viennent avec de nouvelles exigences, et un nouveau mode d'apprentissage. Ils apprennent plus hors du cadre formel, et mènent toute une vie sur les structures sociales. Les nouvelles aptitudes des apprenants 2.0 acquises hors du cadre formel constituent une quantité d'informations abondante et un savoir riche duquel il serait judicieux de tirer profit dans le processus d'apprentissage. Dans notre démarche nous tentons à travers l'analyse des activités des apprenants dans le cadre informel particulièrement sur les réseaux sociaux, d'enrichir nos connaissances sur leurs profils et ainsi cerner les intérêts spécifiques et y répondre de manière personnalisée.

2.3.2 La convergence du Web social et du Web sémantique pour soutenir les systèmes d'e-learning

Les réseaux sociaux sont de nouveaux médias et un nouveau contexte, technologique, créatif, interactif et communicant sur le web. Une orientation qui est en train de générer de nouvelles formes de création de dialogue et de consommation de l'information et qui peuvent représenter de véritables atouts pour la formation. La plupart des formations à distance ont en commun le fait que les apprenants sont physiquement

éloignés, ne se rencontrent que rarement et ne rencontrent que rarement leurs enseignants. Or les réseaux sociaux, avec leurs fonctionnalités à vocation relationnelle et sociale, posent l'usager dans un espace qui lui est propre, qui lui ressemble, convivial et entouré de ses amis. Ainsi ils facilitent une forme de communication spontanée, plus ou moins humanisée par les contenus multimédias (images, vidéos, sons, liens...). On y retrouve aussi des critères de sociabilité et de "confort", ainsi que d'activité dense, qui sont absents dans la plupart des plates-formes d'e-learning traditionnels. Cependant avec toute la convivialité offerte par les réseaux sociaux, cela n'exclut pas le bruit énorme au milieu duquel se perd l'information qui devient péniblement exploitable. Il faut donc instaurer de l'ordre et de la structuration dans les approches de catégorisation de contenu adopté sur les sites Web 2.0. La classification des ressources doit se faire non seulement à base d'étiquetage mais aussi en prenant en compte l'aspect sémantique des ressources (Web de données).

Comme déjà mentionné précédemment l'activité du social tagging génère une quantité abondante de données. Elles peuvent être transformées en des informations sémantiques pouvant être exploités dans l'identification des centres d'intérêt des apprenants ainsi que dans le processus de filtrage des ressources pédagogiques pertinentes. En d'autres termes l'enrichissement du profil apprenant avec ses intérêts est possible grâce à l'enrichissent sémantique des hashtags qu'ils utilisent. L'inclusion de l'aspect sémantique aux hashtags donne la possibilité de filtrer les ressources pédagogiques sur le système d'e-learning d'une manière individuelle, et en fonction du profil enrichi de l'apprenant.

Le nouveau e-learning tend à favoriser la personnalisation des apprentissages, dans le but de rendre plus efficace une situation d'apprentissage à distance à destination d'une génération 2.0 exigeante et ayant un savoir-faire non négligeable. Il s'agit d'une approche à double dimension qui vise d'un côté à personnaliser l'information utile et à faciliter son accès à l'apprenant (dimension Web Sémantique), et à introduire les pratiques communautaires de la communication pour une meilleur implication de l'apprenant (dimension Web Social). Nous percevons donc que la personnalisation de l'apprentissage se situe à la croisée des trois dimensions qui ont connu un grand essor lors des dernières années : le Web social, le Web sémantique et les systèmes d'apprentissage. Dans cette vision nous soulignons l'utilité perçue d'articuler les besoins de personnalisation dans les systèmes d'e-learning, avec les pratiques du Web Social tout en prenant en considération la structuration des contenus par les technologies

sémantiques.

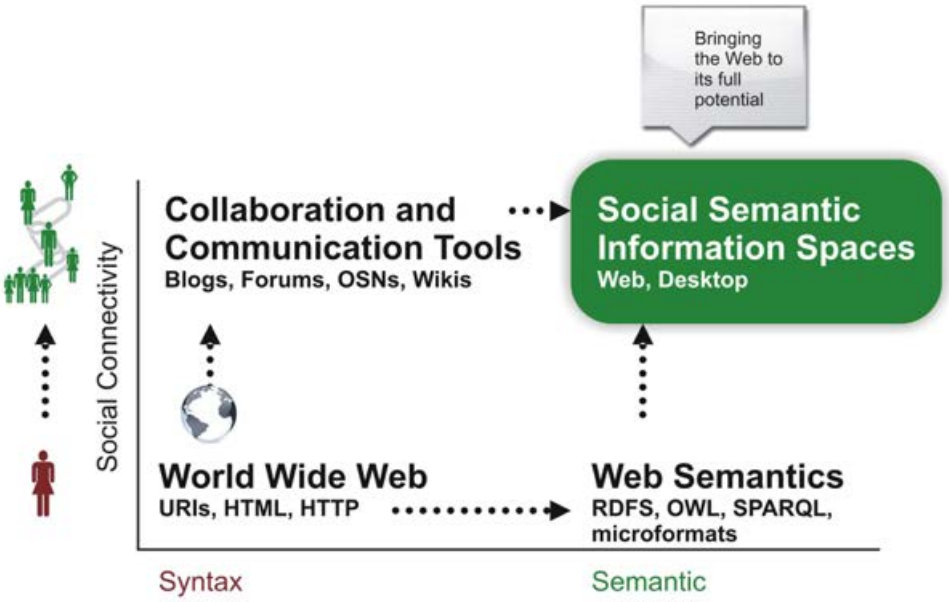

FIGURE *2.12: les espaces d'informations sociales sémantiques* [103]

En résumé, nous constatons qu'une masse de données sociales abondantes est publiée sur différentes structures sociales : facebook, twitter, instagram, foursquare etc. Ces données sont sous des formats hétérogènes, et des identités distribuées ce qui rend l'expérience Web social de moins en moins évidente. Par ailleurs, l'objectif majeur du Web sémantique est de faciliter la compréhension et l'interprétation des informations par la machine, et de permettre une conceptualisation partagée dont le but de faciliter les inférences.

Les plates-formes d'e-learning aujourd'hui ont besoin de supporter différentes approches d'apprentissage telles le formel et l'informel, le personnel et le social à la fois. Le Web social ne pourrait opérer sans des technologies sémantiques. Par conséquent la convergence du Web social et du Web sémantique permet de venir à bout de l'hétérogénéité des données sociales. Les gens seront reliés entre eux, et ceci est d'une grande valeur pour le processus de personnalisation des parcours d'apprentissage. Les recommandations des ressources ou/et des groupes ou personnes susceptibles d'intéresser un apprenant seront plus pertinentes.

La section suivante est consacrée à une étude des différentes approches de recommandation. Nous précisons les avantages et les inconvénients de chacune des approches et nous soulignons les éléments clés nécessaires pour la construction notre approche de recommandation dans un environnement d'e-learning.

2.4 Les systèmes de recommandation

Les systèmes de recommandation ont connu leur essor depuis les années 1990. Selon la définition générale donnée par Robin Burke [104], un système de recommandation est un système capable de fournir des recommandations personnalisées ou permettant de guider l'utilisateur vers des ressources intéressantes ou utiles au sein d'un espace de données important. En d'autres termes les systèmes de recommandation estiment les préférences d'un utilisateur pour lui proposer des ressources personnalisées. Ils ont pour vocation de faciliter le traitement des informations dont le volume et la complexité sont en accroissement continu. Dans cette partie nous commençons d'abord par présenter les différentes approches possibles pour la recommandation, ensuite nous nous focalisons sur les différents travaux effectués dans les systèmes de recommandation pour le e-learning. Nous nous intéressons principalement à la recommandation personnalisée de ressources pédagogique dans un environnement en se basant sur le profil enrichi avec l'ensemble de ses hashtags sémantisés. Ceci dit, que nous nous intéressons également à la modélisation du profil de l'apprenant comme élément clé de la recommandation personnalisée. Plus de détails ainsi que les travaux autour de ces points seront donnés dans la section 2.5.

2.4.1 Les types de systèmes de recommandation

Il existe différents types de systèmes de recommandation selon le type de données à recommander, selon les informations nécessaires pour la recommandation disponible et aussi en fonction des objectifs visés. Mais la plupart des travaux résument les systèmes de recommandations dans les grandes approches suivantes : la recommandation à base du contenu [105] et la recommandation sociale (filtrage collaboratif) [106] qui sont combinées ou réutilisées dans d'autres approches telles que la recommandation hybride [107] ou encore personnalisée [108, 109].

2.4.1.1 Recommandation sociale (filtrage collaboratif)

Ce type de recommandation se réfère aux comportements passés des utilisateurs pour recommander du contenu à des utilisateurs nouveaux dans le système et ayant des intérêts similaires. La recommandation est effectuée alors par les utilisateurs pour d'autres utilisateurs, il s'agit de prédire les ressources qui peuvent intéresser l'utilisateur en analysant les activités des utilisateurs ayant les mêmes comportements que ce dernier. L'hypothèse étant si les utilisateurs ont partagé des choix et des intérêts similaires dans le passé, ils ont de fortes chances d'avoir les mêmes intérêts et de faire le même choix dans le futur [110]. Ce type de recommandations ressemble à celui existant sur Amazon, Youtube ou encore Myspace.

Les algorithmes de recommandation sociale analysent les données du comportement, les activités et les traces des utilisateurs pour tenter de faire de la recommandation à d'autres utilisateurs partageant les mêmes caractéristiques. Ces algorithmes utilisent généralement des mécanismes basés sur le voisinage. Ces techniques de recommandation sociale ou encore appelées filtrage collaboratif [111] peuvent être classées en deux catégories [110] : les méthode du voisinage le plus proche (user-centric / memory-based) et les méthodes à base de corrélation entre contenus dans le voisinage le plus proche (item-centric / model-based) que nous détaillons ci-après.

Méthode du voisinage le plus proche (user-centric / memory-based) [112] Cette approche sélectionne les n profils les plus similaires au profil de l'utilisateur cible de la recommandation, en se basant sur la similarité de leurs intérêts et préférences. Pour chaque ressource auquelle l'utilisateur ne s'est pas encore intéressé, une prédiction est effectuée à base des notes assignés par les utilisateurs voisins aux différentes ressources. L'inconvénient majeur de cette méthode est qu'elle suppose deux conditions essentielles à savoir :

- les utilisateurs qui ont eu des goûts similaires dans le passé auront des goûts similaires dans le futur.
- Les préférences des utilisateurs restent stables et cohérentes dans le temps.

Plus concrètement le filtrage collaboratif à base des notes est tel que figure 2.13 :

Ces notes attribuées par les utilisateurs aux différentes ressources prennent la forme d'une matrice de deux dimensions dont chacune des cases représente l'avis (rating) donné par un utilisateur pour une ressource. Pour recommander à un utili-

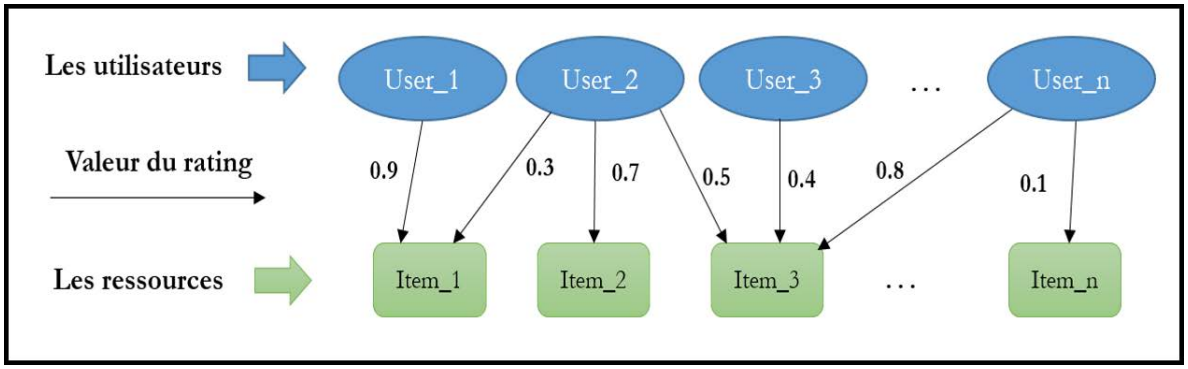

FIGURE *2.13: Filtrage collaboratif à base des notes des utilisateurs*

sateur une ressource sur laquelle il n'a pas d'avis on compare ses ratings avec les autres utilisateurs sélectionnés. Ainsi la prédiction est calculée non seulement à base de l'utilisateur le plus semblable (user centric) mais également à partir de la moyenne pondérée de plusieurs utilisateurs. Le poids donné à la notation de chaque utilisateur est déterminé par le degré de corrélation entre cet utilisateur et l'utilisateur pour qui on désire faire la recommandation. Plusieurs mesures de calcul de similarité existent comme la corrélation de Ringo, ou le coefficient de corrélation de Spearman, mais la mesure la plus utilisée est le coefficient de corrélation de Pearson [113].

Les recommandations doivent être faites à partir de notes données par un grand nombre d'utilisateurs ce qui peut impacter significativement la performance des algorithmes quand le nombre d'utilisateurs atteint une certaine limite. Ceci fait qu'une fois un certain seuil dépassé une sélection des meilleurs voisins est nécessaire pour déterminer les voisins les plus pertinents pour la sélection. L'algorithme le plus souvent utilisé dans ce genre de cas est le k-nearest neighbor(k-NN) [114]. Il permet de sélectionner uniquement les k-meilleurs voisins ayant la valeur de corrélation la plus élevée. Ou encore sélectionner seulement les voisins ayant une corrélation qui dépasse un certain seuil (correlation-thresholding) [115]. Cependant la limitation de ces méthodes réside dans le fait que la bonne valeur du seuil est fortement liée au nombre d'utilisateurs ce qui impacte les prédictions.

Les recommandations ne se limitent pas juste aux ratings ou bien aux like/dislike, l'approche peut aussi prendre en compte d'autres éléments implicites en analysant les comportements d'un utilisateur, ses avis, ou encore ses goûts pour lui proposer d'autres suggestions. Cependant la recommandation sociale (user-centric) montre des limites lorsqu'on est confronté à un nombre important des utilisateurs voisins et aussi

des ressources à recommander. La recommandation temps réel devient très difficile . Alors pour contourner le problème, une autre technique est implémentée dans ce genre de cas, la recommandation sociale item-centric (à ne pas confondre avec la recommandation objet).

Méthode à base de corrélation entre contenus dans le voisinage le plus proche (item-centric / model-based) [116] Dans la recommandation sociale item-centric, au lieu de mesurer la corrélation entre les utilisateurs les ratings sont employés pour mesurer la corrélation entre les contenus. En d'autres termes cette approche se base en premier lieu sur la recherche des contenus similaires et ensuite elle fait de la recommandation à l'utilisateur. Un traitement préalable est effectué sur la matrice de ratings pour sélectionner les contenus similaires et ainsi réaliser des recommandations temps réel contrairement l'approche user-centric très gourmande en terme de ressources mémoire. La similarité entre les ressources est calculée via l'indicateur de similarité cosinus ajustée, dont les valeurs possibles varient dans l'intervalle [-1,1]. Ainsi la prédiction d'une notation (rating) pour un item peut être faite à la base des notations (ratings) effectués par l'utilisateur sur les items similaires.

L'idée de base de cette approche est donc de calculer à l'avance et en temps réel la matrice de similarité entre ressources, et ensuite prédire la recommandation d'une ressource pour un utilisateur. En calculant à la fois les ressources similaires et la valeur moyenne des ratings de ces ressources données par les utilisateurs voisins.

Avantage et inconvénients d'une recommandation sociale

Avantages :

- Ne pas avoir besoin de connaissances préalables sur les propriétés intrinsèques de la ressource.
- La recommandation repose sur du déclaratif (notes, commentaires) explicite fourni par les utilisateurs.

Inconvénients :

- **Scalabilité :** Plus le nombre de ressources et utilisateurs du système est important plus cela demande une puissance dans les calculs et ainsi beaucoup de ressources mémoire, vu qu'il y a un besoin permanent de garder des données

en mémoire pour générer des recommandations pour l'approche user-centric au dépit de sa précision.

- **Le démarrage à froid :** Ce type de systèmes a besoin d'un nombre important de données et d'utilisateurs pour répondre efficacement. Alors le lancement d'une requête de recommandation à nombre réduit de données peut impacter la qualité de cette dernière.
- **Rareté :** La tâche de corrélation peut s'avérer très difficile dans certains cas où même si le nombre de ressources du système est important très peu d'entre elles ont été notées par les utilisateurs. Ce qui veut dire que très peu de ressources suscitant un intérêt commun des utilisateurs.

2.4.1.2 Recommandation objet (filtrage à base du contenu)

Ce type de recommandation vise à recommander du contenu (des objets) à base d'une analyse des descriptions et caractéristiques des éléments précédemment notés par un utilisateur, et de construire un modèle ou un profil des intérêts des utilisateurs sur la base des caractéristiques des objets notés par l'utilisateur [117]. Un profil est créé pour chaque objet à base de ces caractéristiques. Par exemple [110] pour décider de recommander un livre à utilisateur nous nous basons sur les caractéristiques de ce dernier, son genre, le sujet évoqué, son auteur, son éditeur etc. Ceci permettra alors de proposer à l'utilisateur un livre susceptible de l'intéresser et qui soit du même genre ou évoquant un sujet qui s'approche ou genre qu'il a l'habitude de lire. Le processus de recommandation consiste essentiellement à faire correspondre les attributs de profil de l'utilisateur avec les attributs du contenu.

Les algorithmes de ce type de recommandation se focalisent sur la construction de modèles afin de trouver des éléments semblables entre les données. Ils évaluent le degré de similarité entre un contenu pas encore vu par un utilisateur et ces contenus positivement évalués dans le passé. Les algorithmes de recommandation objet attribuent des poids aux attributs d'un objet. Le poids dépend de l'importance de l'attribut à l'utilisateur (plus l'attribut est important plus le poids est élevé), ainsi à base de ces poids la recommandation est effectuée. Un autre algorithme souvent utilisé dans le cas de la recommandation objet est la méthode de retour de pertinence de Rocchio [118] (Rocchio's relevance feedback) ou classification de Rocchio. Cette méthode prend en considération l'avis des utilisateurs sur les recommandations proposées, son principe est simple il suppose que l'utilisateur étant le seul à savoir pertinemment ce qu'il

cherche comme résultat, il serait capable de juger alors les résultats retournées par le système de recommandation. Partant de cette hypothèse les systèmes exploitent les jugements des utilisateurs pour affiner le résultat des recommandations. Avec cette méthode la requête initiale de l'utilisateur est affinée au fur et à mesure que ce dernier retourne ses jugements de pertinence sur les ressources consultés. Les recommandations sont ainsi reformulées en fonction des retours des utilisateurs pour plus de pertinence.

Avantage et inconvénients d'une recommandation Objet [117]

Avantages :

- Contrairement à la recommandation sociale ce type de recommandation ne nécessité pas un grand nombre d'utilisateurs.
- Les recommandations peuvent êtres générée même si on ne dispose que d'un seul utilisateur.

Inconvénients :

- Un taux de subjectivité élevé principalement dans des propriétés d'ordre qualitatif, étant donné que les attributs d'un objet sont généralement renseignés par les propriétaires du contenu.
- Possibilité d'erreur très élevée, vu l'attribution manuelle des attributs par les utilisateurs.

2.4.1.3 Recommandation hybride

Comme son nom l'indique ce type de recommandation combine plusieurs approches de recommandation. Son but est de résoudre les problèmes engendrés par chacune des approches utilisées séparément, à savoir la rareté et le démarrage à froid. Le but de cette combinaison est d'améliorer la qualité des prédictions. Ainsi par exemple l'algorithme de recommandation objet peut être utilisé en premier lieu pour fournir les propriétés des contenus, ensuite un algorithme de recommandation sociale collaboratif peut aider à résoudre les limitations. Quand on est face à un problème de rareté (les utilisateurs n'ont pas évalué assez d'objets) lors d'un filtrage collaboratif, l'approche objet peut résoudre cette lacune via les caractéristiques du contenu.

Parmi les systèmes ayant recours à une approche hybride pour réaliser des recommandations nous citons Amazon et Google [119]. Amazon par exemple combine trois approches de recommandation. Amazon prend en considération le comportement individuel passé de l'utilisateur (recommandation personnalisée) d'une part, les caractéristiques du contenu d'autre part (recommandation objet) et également les comportements des autres utilisateurs (recommandation sociale). Google de son côté mixe les trois approches : il utilise la recommandation personnalisée, ainsi les résultats de recherche sont personnalisés en se basant sur des informations telles que la localisation, et l'historique des dernières recherche. Il fait la recommandation sociale dans le sens où il utilise les liens entre les pages Web et aussi les contenus provenant des cercles google+ des utilisateurs. Et il propose également la recommandation objet en utilisant une approche sémantique de son moteur de recherche via la fonction « did you mean ».

Le tableau 2.2 ci-joint établi par [120] résume et fait une classification des différentes techniques de recommandation dans les trois approches.

Tableau 2.2: classification des différentes techniques de recommandation [120]

	Techniques de recommandation communément utilisés	
Approche de Filtrage	**A base des heuristiques**	**A base des modèles**
A base du contenu	- TF-IDF (Recherche d'information) - Clustering	- Classificateur bayésiens - Clustering - Arbres de décisions - Réseaux de neurones artificiels

Collaboratif	- Voisinage le plus proche (cosinus, corrélation) - Clustering - Théorie des graphes	- Réseaux bayésiens - Clustering - Réseaux de neurones artificiels - Régression linéaire - Modèles probabiliste
Hybride	Combinant les deux approches de filtrage collaboratif et à base du contenu utilisant : - Combinaison linéations des ratings (évaluations) prédis - Systèmes de votes (voting schemes)	Combinant les deux approches de filtrage collaboratif et à base du contenu par : - Utilisation du modèle d'une approche comme partie du modèle de l'autre approche. - Construction de modèle unique d'unification.

Comme cité précédemment nous nous intéressons dans le cadre de nos travaux à proposer une architecture de recommandation multidimensionnelle. Nous sommes donc dans le cas d'une recommandation hybride, car notre approche multidimensionnelle inclut à la fois le filtrage sociale (collaboratif), le filtrage à base du profil enrichi (personnalisé), et aussi un filtrage avec des statistiques systèmes issues des interactions d'un apprenant avec la plateforme e-learning (filtrage à base du contenu) car la recommandation dans ce cas se fait en fonction des objets déjà consultés par un apprenant antérieurement. Notre approche théorique est générale, mais nous nous focalisons dans la partie pratique sur la recommandation personnalisée à base du profil enrichi ou nous utilisons des techniques de clustering pour la recommandation des ressources pédagogiques adéquates au profil d'un apprenant donné.

Dans la section suivante nous allons faire le tour d'horizon des travaux en cours dans la recommandation personnalisée de manière générale et dans les environne-

ments d'apprentissage de manière spécifique.

2.4.1.4 Recommandation personnalisée

La recommandation personnalisée est basée sur les comportements des utilisateurs, L'objectif est de proposer à l'utilisateur une sélection de ressources en fonction d'un profil construit de manière dynamique. En d'autres termes le système récupère l'historique des recherches de l'utilisateur et aussi ses préférences. A partir de ces informations le système établit un profil à base duquel les recommandations sont effectuées pour satisfaire de ses besoins [121]. Dans ce type de recommandation les chercheurs combinent souvent les approches citées précédemment pour effectuer les recommandations (tableau 2.2).

Depuis le premier moteur de recommandation personnalisée né en 1992 [122], les systèmes de recommandations personnalisées ont été largement étudiés et plusieurs travaux ont vu le jour. Nous nous focalisons principalement sur la recommandation personnalisée dans le contexte de l'éducation. Nous donnons plus de détail sur ce type de recommandation ainsi que les travaux cernant cette problématique dans la section suivante.

2.5 La recommandation personnalisée dans un environnement d'e-learning

La nouvelle génération des systèmes d'apprentissage cherche à intégrer de nouvelles approches centrés utilisateurs lui permettant d'être actif et participer à la construction du savoir afin d'affiner la personnalisation des contenus proposés par le système. Etant donné que la personnalisation des recommandations dépend du profil apprenant, nous devons alors avoir assez d'information afin de pouvoir l'enrichir et ainsi procéder à une recommandation personnalisée à base du profil enrichi. Ceci dit qu'il est nécessaire de modéliser le profil avant de procéder à son enrichissement. De ce fait nous étudions d'une part les différentes propositions de modélisation pour déterminer le modèle qui correspond le plus à notre approche. Et d'autre part nous faisons le tour des travaux sur la recommandation personnalisée dans le contexte de l'éducation.

2.5.1 La modélisation du profil apprenant

Le profil apprenant retient une attention particulière dans notre cas d'étude car le but final est de faire de la recommandation du contenu pédagogique à base du profil de l'apprenant. Dans cette partie nous tentons d'explorer les méthodes et standards de modélisation du profil apprenants.

Avant de traiter les modèles de l'apprenant, il est judicieux de définir ce qu'est un modèle apprenant. Diverses définitions existent dans la littérature, Selon [123] le modèle apprenant est défini comme un ensemble d'informations propres à un apprenant, des informations qui portent souvent sur les connaissances et savoir-faire que le système attribue à l'apprenant au vu de son comportement. [124] propose une vision de l'apprenant comme un quadruplet (P, C, T, H) où P décrit le niveau des connaissances procédurales de l'élève (ses savoir-faire), C décrit les connaissances conceptuelles de l'élève (ses savoirs), T décrit les traits particuliers de l'élève comme son caractère et H est l'histoire de l'apprentissage (l'historique des sessions). [125] pense le modèle d'apprenant comme central dans un système d'apprentissage adaptatif puisqu'il reflète la compréhension de l'apprenant (et éventuellement d'autres aspects comme ses buts, ses préférences d'apprentissage, ses motivations, etc.). Il permet à un système de s'adapter de façon dynamique aux besoins d'apprentissage de l'utilisateur. Dans la définition de [126] les auteurs soulignent l'aspect dynamique du modèle. Ainsi, ils définissent le modèle de l'apprenant comme étant un portrait des connaissances de l'élève qui s'enrichit à chaque étape de l'apprentissage. Ils considèrent également que les paramètres quantitatifs ne forment pas un modèle de l'apprenant. En d'autres termes le modèle de l'apprenant ne se réduit pas à un ensemble de scores (par exemple 4 bonnes réponses sur 10), mais se compose d'un ensemble d'informations sur les méthodes qu'il a appliquées, les causes communes à toutes ses erreurs, etc.

D'après [127], les informations du profil d'un apprenant doivent êtres décrites de manière rigoureuse, ainsi le modèle décrivant le profil doit être très bien décrit pour faciliter sa réutilisation par les différentes plates formes e-learning avec lesquelles l'apprenant est en interaction. Ceci dit que la modélisation du profil selon les standards est très intéressante pour la normalisation et la facilité des échanges. Dans ce contexte nous décrivons les standards les plus importants et les plus connus pour la modélisation d'un profil d'apprentissage à savoir PAPI [128], IMS LIP [129] et IMS RDCEO [130], et IMS e-Portfolio [131].

PAPI (Public and Private Information) : est un standard proposé par le groupe Learner Model Working Group de l'IEEE, ce Standard spécifie à la fois la syntaxe et la sémantique du modèle de l'apprenant. Il décrit les décrit un apprenant selon ses connaissances, ses compétences, ses performances et également ses relations avec des acteurs du systèmes. PAPI est l'un des premiers standards de modélisation des informations sur les apprenants, il structure ses informations en six catégories :

- **Informations personnelles :** informations personnels sur l'apprenant (nom, adresse, etc.) ;
- **Relations :** informations sur les relations entretenues avec les autres apprenants (camarade, professeur, etc.) ;
- **Sécurité :** informations sur la sécurité (clés publiques et privées, mot de passe, etc.) ;
- **Préférences :** Destinée à améliorer l'interaction homme/machine et l'adaptation automatique des systèmes aux besoins spécifiques de l'apprenant (la langue d'apprentissage, le style d'apprentissage, etc.) ;
- **Performances :** histoire de l'apprenant, son travail en cours, et ses objectifs futurs (notes d'apprentissage, rapports, etc.) ;
- **Portfolio :** Collection des travaux illustratifs de ses capacités (travaux et expériences précédentes, etc.).

Le limitation majeur de ce standard c'est son incapabilité de prendre en considération plusieurs informations de l'apprenant pouvant être échangées entre divers systèmes d'apprentissage. ce qui explique son extension vers IMS-LIP par IMS [132].

IMS-LIP (IMS Learner Information Package) : est une spécification pour la structuration des informations sur les apprenants. Il se base sur PAPI mais il est plus riche en terme d'informations et aussi plus général. Ce standard se concentre principalement sur l'historique et l'expérience d'apprentissage d'un apprenant. Son but est de faciliter l'échange des informations sur les apprenants entre systèmes éducatifs, systèmes de gestion d'apprentissage, etc. IMS-LIP est structuré en onze catégories à savoir : l'identification, le but, les qualifications, certifications & licences (QCL), l'activité, les intérêts, les Compétences, la transcription, l'affiliation, l'accessibilité et la sécurité. (nous traiterons plus en détail ces catégories lors de l'enrichissent du profil dans le chapitre 4.5).

IMS-RDCEO (IMS Reusable Definition of Competency or Educational Objectives) : est une spécification [133] définissant un modèle d'information qui à la fois décrit, met en référence et échange des définitions des compétences, principalement dans le contexte de la formation en ligne et l'apprentissage distribué. IMS-RDCEO permet l'interopérabilité entre les systèmes qui traitent des informations de compétence en leur fournissant un moyen de se référer à des définitions communes avec des significations communes [134]. IMS-RDCEO emploi le mot compétence dans un sens très général qui résume les qualifications, les tâches, et même les résultats d'apprentissages. Les informations de la compétence sont résumé en cinq catégories dont seul l'identifiant et le titre sont obligatoire [130] :

- **Identifiant :**un identifiant unique de la compétence ;
- **Titre :**Il s'agit d'une description textuelle de la compétence ;
- **Description :** une description (optionnelle) de la compétence ;
- **Définition :** Une description facultative structurée de la compétence qui est souvent décrite en termes d'étapes, de critères, d'indicateurs, de capacités de production, d'habiletés, de niveaux. Cet élément est structuré grâce à la notion de « Statement ».
- **Statements :** Ils sont décrits par un identifiant, un nom, un texte de description et des définitions de mots utilisés dans la description.

IMS e-portfolio [131] : est une collection de renseignements personnels au sujet d'un apprenant, elle représente ses réalisations, ses objectifs, ses expériences et autres dossiers personnalisés que l'apprenant peut présenter à l'école, aux employeurs, ou à d'autres entités. Les informations, les objets de la performance, ainsi que les réalisations d'une personne, comme enregistré dans un e-Portfolio opèrent à travers les institutions et les pays tout au long de la vie. en d'autres termes c'est une forme numérique pour un apprentissage à vie et dans des contextes variés(académique, professionnel, personnel). La spécification ePortfolio :

- Prend en charge de la promotion de l'apprentissage continu pour de nombreuses initiatives gouvernementales
- Rend plus faciles l'échange du Portfolio de l'école au travail
- Permet aux éducateurs et aux institutions de mieux suivre les compétences
- Améliore l'expérience d'apprentissage et améliore le perfectionnement des employés

D'autres travaux récents optent pour la modélisation selon des ontologies, étant donné que l'ontologie constitue une solution pour la représentation et le partage des connaissances de manière standardisé. Nous citons le travail de [135] où les auteurs ont opté pour la modélisation d'un apprenant dans un système hypermédia selon une ontologie. L'ontologie apprenant est basée sur des résultats des travaux des théories cognitives pour la description des profils apprenants. Les auteurs de [136] qui se sont plutôt concentrés sur la modélisation des connaissances de l'apprenant sous forme d'une ontologie nommée "Diag-K". Cette ontologie contient toutes les connaissances de l'apprenant permettant d'analyser son niveau par rapport à un thème enseigné sur le EIAH, et ainsi élaborer automatiquement son profil.

Pour conclure, nous constatons que les standards cités précédemment contiennent des informations importantes pour la représentation d'un apprenant de manière standardisée. Nous avons opté pour le IMS-LIP étant donné qu'il a pu devancer PAPI en approchant le plus possible l'apprenant et en proposant une meilleure description de ses informations. Et également car il est générique contrairement à IMS-RDCEO qui se focalise principalement sur la description des compétences. Le IMS-ePortfolio est égalment très intéressant pour une modélisation des apprenants à vis même après qu'ils quittent leur institution académique vers le milieu professionnel. Mais dans notre approche nous nous contentons du suivi de l'apprenant uniquement lors de sa formation. IMS-LIP nous permis d'approcher notre objectifs de modélisation et par la suite celui de l'enrichissement de profil apprenant après modélisation.

2.5.2 Travaux autour de la recommandation personnalisée

Le but de la recommandation dans un environnement d'apprentissage est d'aider l'apprenant et de faciliter son choix des ressources utiles et intéressantes dans sa formation. Dans cette perspective un nombre important de systèmes de recommandation est proposé dans le domaine de l'éducation que ce soit dans le contexte formel ou informel. Nous commençons par citer [137] qui s'attaquent au problème de surcharge d'information dans le cadre de l'apprentissage à distance. Ils proposent un processus en quatre étapes pour la recommandation d'une sélection de ressources intéressantes dans une grande collection d'objets d'apprentissage. Ils montrent dans leurs résultats que la majorité des utilisateurs n'ont pas le temps ni la volonté de parcourir toutes les ressources proposées par le système et c'est pourquoi ils trouvent l'aide dans le choix des ressources utiles très intéressant. Les auteurs [138] où les auteurs basent leur

approche sur un système à base de trace déjà existants, ils définissent deux modèle un définissant les profils et l'autre les ressources du système. Ainsi ils proposent une implémentation des services de personnalisation au sein de l'outil ARIADNE Finder dont les fonctionnalités concernent le filtrage des objets pédagogiques par rapport à la langue et au format de fichiers préférés de l'utilisateur, ainsi que par rapport aux ressources consultées par l'utilisateur.

Toujours dans le même sens les auteurs de [139] présentent une approche de recommandation personnalisée de ressources pédagogiques dans une Communauté de Pratique de E-learning (CoPE). Ils se basent sur les objectifs et les besoins de ces membres ainsi que sur leur feedback sur les différentes ressources existantes dans la mémoire de la CoPE pour la prédiction de la liste de recommandation. Ainsi les stratégies de recommandations sont principalement du filtrage à base du contenu et filtrage collaboratif fusionnés ou utilisés séparément en fonction de la situation rencontrée (nouvel utilisateur, volonté d'apprentissage et/ou de spécialisation). Un prototype de l'approche a été développé et aussi des évaluations qualitatives à base de questionnaires ont été menées sur un comité de pratique e-learning dans le domaine universitaire. Ces évaluations montrent clairement l'intérêt de l'approche pour ces membres. Dans [140] les auteurs se sont intéressés à l'utilisation du filtrage collaboratif dans une plateforme e-Learning pour la recommandation des objets pédagogiques à un utilisateur actif. Ils utilisent une base de données des évaluations des utilisateurs pour identifier le voisinage le plus proche de l'utilisateur actif et prédire ses évaluations pour lui recommander des ressources pertinentes. Dans ce travail plusieurs mesures de corrélation sont exploitées dans le filtrage collaboratif à savoir la mesure de Pearson.

D'autres contributions partent du principe que les systèmes de recommandations personnalisées considèrent généralement les préférences, les intérêts et les comportements de navigations des apprenants pour offrir un service personnalisé. Cependant ils négligent souvent la capacité et le niveau des apprenants dans le processus de recommandation. Dans [141] les auteurs d'un système de e-learning personnalisé basé sur la théorie de réponse à l'item (PEL-IRT) qui considère à la fois la difficulté du matériel pédagogique et la capacité de l'apprenant pour fournir des chemins d'apprentissage individuels pour les apprenants. Ils proposent également une approche collaborative de vote pour mieux ajuster le paramètre de difficulté d'un cours. Les résultats des expériences montrent clairement que l'application de la théorie IRT pour l'apprentissage à base du Web réalise la personnalisation des apprentissages et ainsi aident

les apprenants à apprendre efficacement. Dans [142] les auteurs écrivent un module de recommandation d'un système de tutorat qui s'adapte automatiquement aux intérêts et aux niveaux des apprenants. Ce système reconnaît différents modèles de style d'apprentissage et les habitudes des apprenants à travers leurs logs de connexion, et fait des recommandations personnalisées du contenu pédagogique selon les évaluations de ces séquences fréquentes, fournis par le système Protus. Les expériences ont été réalisées sur deux groupes des apprenants : le groupe de contrôle (ceux ayant appris d'une manière normale et n'ont pas reçu de recommandation ou des conseils dans le cours). Et le groupe expérimental (ayant appris en utilisant le système Protus). Les résultats ont montré la pertinence de l'utilisation du module de recommandation dans la démarche apprenante.

Des travaux récents comme [143] où les auteurs suivent une approche à base des informations présentes sur les réseaux sociaux pour recommander des ressources pédagogiques dans le domaine des EIAH. L'objectif de ce travail est d'intégrer le concept du Learning Management System dans un réseau social en reliant leurs éléments communs avec une approche à base de systèmes de recommandation. Cette intégration se présente sous la forme d'une surcouche dans les réseaux sociaux via l'utilisation d'un plugin WordPress [144]. Toujours dans le cadre des systèmes hypermédia adaptatifs une autre contribution [145] propose une manière innovante de suivre l'activité de l'utilisateur en termes de thèmes d'intérêt par rapport à sa navigation conceptuelle. L'idée de l'approche est de modéliser les ressources du système que ce soit les utilisateurs, les documents, ou les services offerts par rapport à une ontologie unique. De cette façon les annotations associées aux ressources visualisées renvoient l'information sur les intérêts de l'utilisateur et ainsi le système construit implicitement son profil. Les auteurs ont opté pour une recommandation personnalisée à base du contenu et dont l'algorithme prend en compte la hiérarchie des classes caractérisant l'ontologie et la similarité entre les concepts ontologiques. Le travail illustre bien l'utilité des ontologies comme support sémantique dans l'amélioration des fonctionnalités de recommandations.

2.6 Synthèse et conclusion

Les plates-formes d'apprentissages e-learning offrent aux apprenants une panoplie de ressources pédagogiques et de technologies de communication et d'échanges

dans un environnement structuré et convivial. Cependant malgré toute la convivialité offerte par les systèmes d'apprentissage à distance il leur faut plus pour susciter l'intérêt des apprenants et éviter leur abandon des formations. Ceci dit que pour motiver un apprenant il faut lui faire gagner du temps en lui proposant des ressources qui répondent efficacement à ses besoins et dans un temps opportun. Aujourd'hui le problème de manque de ressources ne se posent plus, au contraire nous avons une quantité abondante de ressources et le challenge est de faire choisir les plus pertinentes pour un apprenant donné en fonction de son profil.

Les travaux de recherche menés dans le cadre de l'apprentissage à distance et les environnements e-learning proposent tous des méthodes et outils de gestion de connaissance centré utilisateur et visent à mettre en valeur la dimension de personnalisation et ce en tirant profit des technologies récentes à savoir celle du Web social et du Web sémantique. Car d'un côté l'aspect collaboratif du Web social contribue efficacement à améliorer et étendre les fonctionnalités des plateformes d'e-learning. Et l'aspect sémantique du Web 3.0 permettra une meilleur structuration des informations et ainsi la possibilité de combiner divers informations hétérogènes en provenance de différents sources. Ainsi le système pourra améliorer la compréhension et la formalisation des besoin spécifiques et ainsi perfectionner les recommandations proposées.

La plupart des travaux autour de la recommandation dans le cadre des environnements e-learning vu dans la section précédente génèrent les informations sur les profils d'apprentissage à partir des informations que les apprenants doivent renseigner des des formulaires ou à partir des questionnaires et également à partir de traitement des interactions avec le système. Dans notre approche nous proposons d'analyser les traces et les données partagées par un apprenant au sein des réseaux sociaux. Cela nous permettra une meilleur visibilité sur des intérêts et ainsi cibler les ressources qui lui seront pertinent au sein de la plate-forme e-learning.

Notre objectif est d'enrichir les centres d'intérêts d'un apprenant à base des activités recueillies sur les réseaux sociaux. Nous nous intéressons à l'analyse des activités et traces de l'apprenant dont le but est de remédier au manque de rétroaction entre l'apprenant et l'environnement d'apprentissage. Dans cette perspective nous avons mené une étude approfondie de l'état de l'art concernant la modélisation des profils des apprenants au sein d'un LMS et nous avons opté pour le choix d'une modélisation qui se base sur un standard afin d'assurer la facilité de réutilisation des traces recueillies dans et entre les différentes plateformes pour une approche de recomman-

dation personnalisée des ressources pédagogiques dans un cadre général.

Dans ce chapitre nous avons examiner les travaux pionniers dans le sens d'amélioration des systèmes e-learning lorsque le Web Social et le Web sémantique sont impliqués dans l'approche et également les travaux autour de la recommandation dans les environnements e-learning. Dans les chapitres suivants nous présentons en détail notre méthode d'analyse des connaissances sociales pour l'enrichissent du profil de l'apprenant à travers ses activités sociales d'une part. D'autre part nous abordons notre approche de recommandation des ressources pédagogiques à l'apprenant au sein d'une plateforme d'e-learning.

CHAPITRE 3

METHODE D'ANALYSE DES CONNAISSANCES SOCIALES POUR L'ENRICHISSEMENT DU PROFIL DE L'APPRENANT

Sommaire

Nous avons conclu dans le chapitre précédent que l'apprentissage n'est pas seulement un processus de transmission de l'information. C'est surtout un processus d'enrichissement de profils des apprenants à travers leurs activités en dehors de l'environnement d'apprentissage, pour répondre efficacement à leurs besoins spécifiques. Dans ce chapitre nous détaillons notre approche d'extraction et d'analyse des activités sur les réseaux sociaux. Nous construisons un dictionnaire communautaire permettant de désambiguïser les hashtags et les rendre exploitable pour des fins d'enrichissement de profil apprenant sur un environnement d'apprentissage.

3.1 Représentation sémantique des activités et des contributions des utilisateurs sur les réseaux sociaux

L'amélioration de l'expérience d'apprentissage au sein des environnements d'e-learning est très liée à la personnalisation. Afin de proposer des ressources de manière personnalisée à un apprenant, il est nécessaire d'avoir une bonne connaissance de son profil. Le problème de personnalisation des apprentissages revient donc à un problème d'enrichissement du profil apprenant à partir de ses activités et données partagées notamment au sein des réseaux sociaux. L'objectif de ce travail est d'utiliser les informations que l'on peut chercher et déduire des écrits et données partagées des apprenants sur les réseaux sociaux en tirant profit des technologies sous-jacentes notamment le Web Sémantiques, Web Social. Pour améliorer la qualité d'apprentissage à distance. il s'agit d'enrichir le profil d'un apprenant à partir des connaissances recueillies sur ses contributions sur divers médias sociaux pour une recommandation personnalisée. Plus précisément, nous procédons à l'extraction des contributions des apprenants sur les réseaux sociaux, pour ensuite les analyser et les structurer de façon à les rendre exploitables à des fins de recommandation des contenus pédagogiques lors d'une démarche d'apprentissage sur la plate-forme d'e-learning.

Les écrits et les données des utilisateurs sur les réseaux sociaux sont en croissance exponentielle au fil du temps. Afin de lier et retrouver facilement ce qu'ils produisent au sein de cette grande masse de données, les utilisateurs étiquettent les

ressources en utilisant les hashtags. Les hashtags sont devenus une solution légère pour classer et rechercher des informations sur le Web et aussi une solution très conviviale pour l'utilisateur final. Malheureusement, malgré tous leurs points forts les hashtags ne sont pas automatiquement exploitables. Les hashtags sont ambigus et ne peuvent pas être définis clairement. Ils sont créés par les utilisateurs pour étiqueter et rechercher du contenu, et ils sont souvent des néologismes, abréviations ou concaténations de mots. Il est donc difficile de les exploiter à l'état brut. Il faut les désambiguïser en prenant en compte l'aspect sémantique. Dans nos travaux, nous avons conçu un processus permettant de désambiguïser les hashtags en se basant sur les définitions fournies par les utilisateurs sur divers services en ligne (tagdef.com, hashtags.org) afin de générer un dictionnaire de hashtags qui organise les définitions de chaque hashtag en unités de sens. Ce dictionnaire fait l'objet de la section 3.2.2, ceci nous permet d'enrichir les centres d'intérêts des profils apprenants et ainsi de faire des recommandations personnalisées.

3.2 Folksionary

Nous nous concentrons sur l'étude des hashtags écrits par les personnes au sein de leurs messages. Les hashtags sont auto-émergents (n'importe qui peut créer n'importe quel hashtag) et mal définis (on trouve des définitions manuelles répertoriées dans des dictionnaires en ligne, parfois contradictoires, avec des homonymes). Nous avons donc voulu construire en première instance un dictionnaire consolidé de hashtags, qui à la manière d'un dictionnaire classique, associe à chaque terme (hashtag) un ensemble de sens, chaque sens pouvant avoir plusieurs définitions. Ce dictionnaire est appelé Folksionary un mot-valise combinant les mots folk (le peuple, les gens) et dictionary (dictionnaire), signifiant donc un "dictionnaire du peuple". Les outils du Web sémantique et des techniques de data mining ont été utilisés pour arriver à ces fins. Nous avons proposé et implémenté une approche permettant de lever l'ambiguïté des hashtags en se basant sur les définitions. Pour cela nous calculons la similarité sémantique entre toutes les définitions d'un hashtag donné pour ensuite les regrouper en unités de sens grâce à un algorithme de clustering, et finalement générer une version intelligible du Folksionary pour l'humain à l'image d'un dictionnaire classique.

Dans cette section, nous détaillons notre approche permettant de regrouper en différents sens les définitions rédigées par les utilisateurs sur les services en ligne

pour un hashtags donné.

3.2.1 Formalisation mathématique

Nous traitons le problème suivant : étant donné un dictionnaire généré par les utilisateurs, dans lequel l'utilisateur peut entrer ses propres définitions pour un hashtag donné, comment le structurer de façon similaire à un dictionnaire classique. Plus précisément, pour chaque hashtag du dictionnaire, comment regrouper ses définitions en unités de sens ? Nous ramenons ce problème à un problème de mesure de similarité entre les définitions d'un hashtag.

Nous formalisons le problème en utilisant une fonction de similarité **simMeaning** permettant de calculer la similarité sémantique paire à paire de l'ensemble des définitions d'un hashtag. **simMeaning** est définie comme suit :

Définition 7 *Si deux définitions def_1 et def_2 pour un hashtag w partagent le même sens la fonction de similarité sera défini comme : simMeaning($def_1(w)$; $def_2(w)$) = 1, sinon : simMeaning($def_1(w)$; $def_2(w)$)= 0.*

Notre objectif est de partitionner l'ensemble des définitions pour un hashtag donné en des sous-ensembles, tel que dans chaque sous ensemble la similitude entre les paires de ses définitions est égale à 1. Avant de définir la fonction de partitionnement automatique des définitions d'un hashtag nous commençons par définir les concepts suivant :

- W l'ensemble des hashtags,
- $D(w)$ l'ensemble des définitions pour chaque hashtag $w \in W$,
- $S(w)$ l'ensemble des sens possibles pour chaque hashtag $w \in W$, tel que : $S(w) = D(w)$,
- On note aussi $d_{w,i}$ la $i^{ème}$ définition du hashtag $w \in W$, où $i \in [1, |D(w)|]$.

Définition 8 *Nous définissons la fonction de clustering notée $\mathcal{E}$ qui relie chaque définition d'un hashtag $w \in W$ à un sens $s \in S(w)$ du même hashtag comme suit :*

$$\begin{aligned} \mathcal{E} : \quad & D(w) \longrightarrow S(w) \\ & \forall d \in D(w), \exists s \in S(w),\ \mathcal{E}(d) = s \end{aligned} \tag{3.1}$$

Propriété 1 *$\mathcal{E}$ est une fonction surjective (c.f. Equation 3.1), ce qui veut dire que chaque sens $s \in S(w)$ de chaque hashtag $w \in W$ dans le folksionary est constitué d'au*

moins une définition $d_{w,i} \in D(w)$ générée par l'utilisateur sur un service de définition en ligne.

Propriété 2 *$S(w)$ est une partition de $D(w)$, tel que toute définition $d_{w,i} \in D(w)$ générée par l'utilisateur pour un hashtag $w \in W$ appartienne à exactement un sens $s \in S(w)$, ce qui signifie :*

$$\begin{cases} \bigcup s_{w,i} \in S(w) = D(w) \\ \forall s_1, s_2 \in S(w), s_1 \neq s_2 \Rightarrow s_1 \cap s_2 = \emptyset \end{cases} \tag{3.2}$$

La fonction de clustering $\mathcal{E}$ est maximisée quand toutes les définitions avec des significations non similaires sont séparées dans des partitions différentes. Chaque groupe définitions (partition) est considéré comme un sens distinct pour ce hashtag.

Les distances de similarité calculées entre les différentes paires de définitions $def_{w,i} \in D(w)$ pour un hashtag $w \in W$ sont formalisées avec une matrice normalisée appelée matrice de similarité $Dist(w)$, nous la formalisons comme suit :

$$Dist(w) = (dist(def_{w,i}, def_{w,j}))_{1 \leq i,j \leq |D(w)|} \tag{3.3}$$

Exemple Dans le but d'illustrer les différents concepts précédemment formalisés nous adoptons l'exemple suivant pour le hashtag #acm :

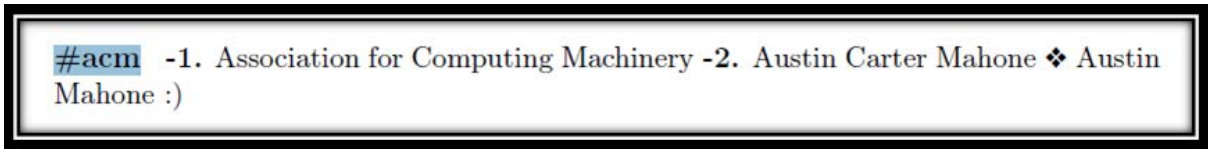

FIGURE *3.1: Le hashtag #acm dans le folksionary*

- $|D(\#acm)| = 3 \Leftrightarrow$ Le hashtag #acm contient trois définitions : def_1="Austin Carter Mahone", def_2="Association for Computing Machinery" et def_3="Austin Mahone :)".
- $|S(\#acm)| = 2 \Leftrightarrow$ La fonction de clsutering $\mathcal{E}$ génère deux sens $s_{\#acm,1}, s_{\#acm,2}$ pour le hashtag #acm.
- Le premier sens $s_{\#acm,1}$ contient une seule définition : def_2, tandis que le second sens $s_{\#acm,2}$ comporte deux def_1 et def_3.

- $s_{\#acm,1} \neq s_{\#acm,2} \Rightarrow ss_{\#acm,1} \cap s_{\#acm,2} = \emptyset$ (une définition appartient exactement à un seul sens).
- La fonction de similarité pour le hashtag #acm est défini comme suit simMeaning($def_1(\#acm)$; $def_3(\#acm)$)=1 pour les définitions qui sont dans la même unité sens, sinon simMeaning($def_1(\#acm)$; $def_2(\#acm)$)=0 et simMeaning($def_2(\#acm)$; $def_3(\#acm)$)=0 pour les définitions ayant des sens différents.

Le but de la formalisation mathématique des différents concepts cités précédemment est de guider le lecteur à comprendre les différentes notations et équations que nous allons par la suite transformer en algorithmes lors de la construction du folksionary (section 3.2.2).

Dans les perspectives de nos travaux nous nous sommes fixés comme objectif de dégager les relations qui puissent exister entre les hashtags du folksionary, plus concrètement les relations existantes entre les unités de sens des hashtags contenus dans le folksionary. Nous avons donc entamer le travail dans ce sens en commençant par la relation de synonymie. Nous définissons ci-après la matrice de similarité entre deux hashtags distincts :

$$\begin{cases} \forall w_1, w_2 \in W, w_1 \neq w_2 \\ Dist(w_1, w_2) = (dist(s_{w,1i}, s_{w,2j}))_{1 \leq i \leq |s(w_1)|, 1 \leq j \leq |s(w_2)|} \end{cases} \tag{3.4}$$

Tel que la similarité entre le i^{eme} sens du hashtag w_1 et le j^{eme} sens du hashtag w_2 est définie comme la moyenne de distance des paires de définitions $def_{w,i}$ et $def_{w,j}$ des deux hahstags w_1 et w_2 :

$$Dist((s_{w,1i}, s_{w,2j}) = \frac{\sum_{k=1}^{|D(w_1)| \times |D(w_2)|} (dist(def_{w,i}, def_{w,j}))_{1 \leq i \leq |D(w_1)|, 1 \leq j \leq |D(w_2)|}}{|D(w_1)| \times |D(w_2)|} \tag{3.5}$$

3.2.2 Processus de construction du folksionary

Pour construire le folksionary, nous effectuons un processus en quatre étapes. Tout d'abord, nous récupérons les définitions des hashtags à partir des services de définitions des hashtags disponibles en ligne (à savoir hashtags.org, tagdef.com). Deuxiè-

mement, pour chaque hashtag, nous procédons à une comparaison paire à paire de ses définitions à travers le calcul de la distance entre les paires de définitions. Lors de la troisième étape, nous appliquons un algorithme de clustering pour chaque hashtag afin de regrouper ses définitions en des groupes de sens. Enfin, nous exportons ces résultats sous la forme d'un document lisible par l'humain ayant une apparence très proche de celle d'un dictionnaire standard. La figure 3.2 illustre cette approche et les sections suivantes donnent le détail de ces quatre étapes.

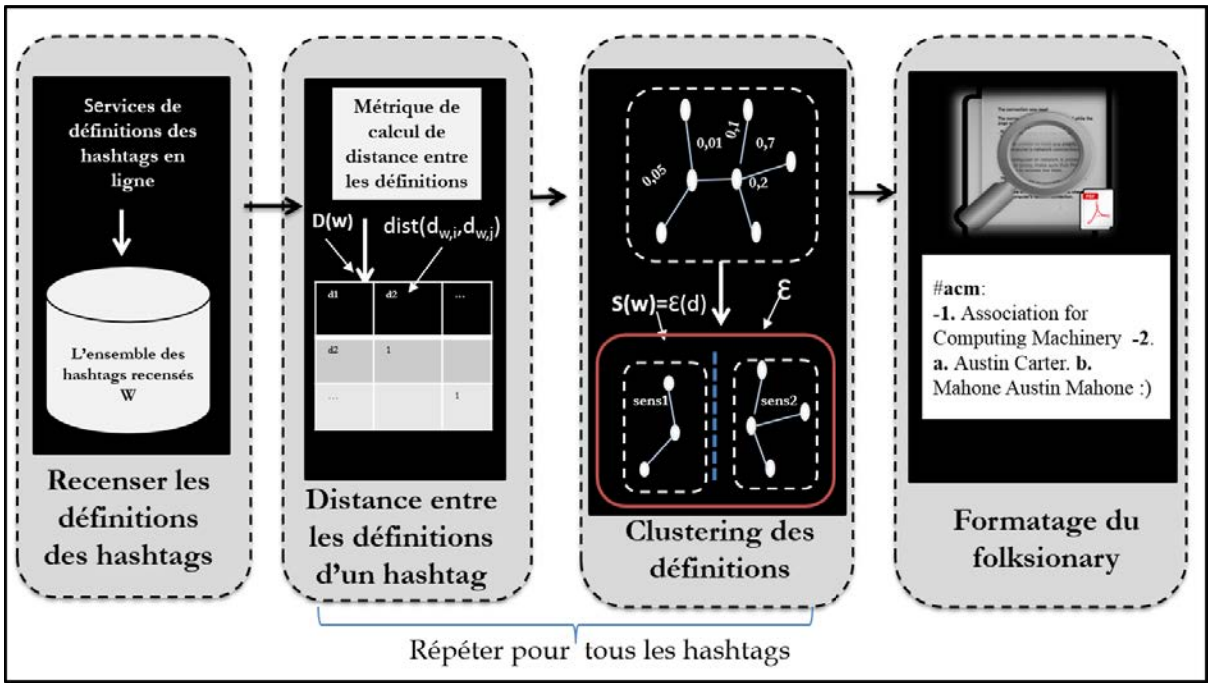

FIGURE *3.2: Approche de construction du folksionary*

3.2.2.1 Recensement des définitions des hashtags

Notre but dans cette étape est de recenser les hashtags les plus utilisés par les usagers des réseaux sociaux ainsi que leur différentes définitions. Autrement dit, nous cherchons à peupler $D(w)$, $\forall w \in W$. Différentes sources de données sur le Web contiennent des définitions de hashtags rédigées par les utilisateurs en langage naturel. Par exemple, *Tagdef.com* et *Hashtags.org* sont des dictionnaires de hashtags en ligne bien connus. Nous recensons les hashtags et les définitions à partir de ces sources. Le processus Web scraping [1] extrait les définitions à partir de chaque page

1. Script d'extraction de données présente sur un site Web.

donnée, et on ne conserve que les définitions rédigées en anglais à l'aide d'un classificateur de langue [2] intensément testé et qui a prouvé son efficacité sur 500 millions de fichiers à travers une variété d'applications industrielles, académiques et de laboratoires gouvernementaux.

Ensuite nous réaliserons le stockage des données recueillies dans la base de données NoSql Cassandra étant donné ces capacités à stocker une grande quantité de données grâce à sa scalabilité horizontale. Ainsi que ses divers avantage par rapport à une base de données classique à savoir : sa tolérance aux pannes du fait que les données d'une instance Cassandra sont automatiquement répliquées vers d'autres instances sur différentes machines. Son riche modèle de données basé sur la notion de clé/valeur qui permet de développer de nombreux cas d'utilisation dans le monde du Web. Son élasticité dû au fait que Le débit d'écriture et de lecture augmente de façon linéaire, chose qui assure la disponibilité du système et évite les interruption au niveau des applications. Sa haute disponibilité, car l'écriture des données est très rapide comparée au monde des bases de données relationnelles. Et également son aspect décentralisé ce qui veut dire qu'il n'y pas de notion de maitre, ni d'esclave, ni de processus qui aurait à sa charge la gestion, ni même de goulet d'étranglement au niveau de la partie réseau. Les données que nous avons recensé sont stockés sous un Schéma unique (figure 3.3).

Nous avons un modèle de données Cassandra à quatre étages :
Keyspace → Column Family → Column List → Column →Data value.

- **Keyspace :** c'est l'équivalent d'une base de données dans les bases données relationnelles. Le keyspace représente l'étage le plus haut dans le modèle de données Cassandra, dans notre cas le keyspace (TagDef ou Hashtags) représente l'ensemble de tous **les hashtags** recensés sur les services en ligne *Tagdef.com* et *Hashtags.org* avec leurs **définitions**.
- **Colonne (Column) :** c'est la plus petite unité de donné dans le modèle de données Cassandra. La colonne est composée d'une clé (le nom de la colonne), une valeur et un instant de création. Dans notre schéma c'est le nom de la colonne c'est **une définition parmi la liste des définitions d'un hashtag** et sa valeur c'est **l'origine d'où a été tiré cette définition**.
- **Ligne (Row) :** c'est une aggrégation de colonnes. La ligne est représentée par une clé et une valeur. Nous distinguons deux types de lignes : les **wide row**

2. https ://tika.apache.org/

permettant de stocker énormément de données, avec beaucoup de colonnes et les **skinny row** permettant de stocker peu de données. Dans notre schéma de données la nom de la ligne est équivalent à un **hashtag de l'ensemble de hahstags** et la valeur c'est **l'ensemble de toutes les définitions du même hashtag**.

– **Famille de colonnes (ColumnFamily) :** c'est l'objet principal de données et peut être assimilé à une table dans le monde des bases de données relationnelles. Toutes les lignes sont regroupées dans une famille de colonnes. La clé de la Famille de colonnes c'est son nom (**Liste des Tags**) tandis que sa valeur est l'ensemble de toutes lignes dans notre schéma **c'est l'ensemble de tous les hashtags avec leurs définitions**

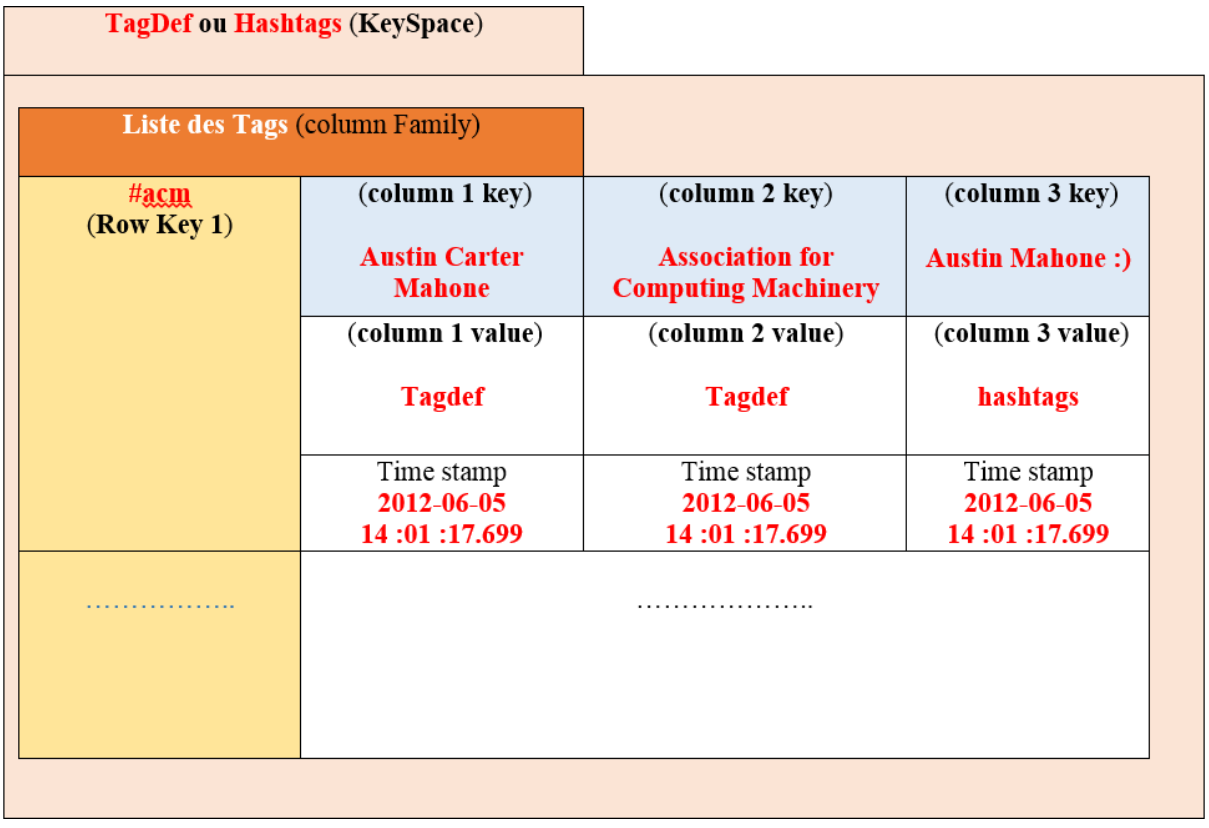

FIGURE *3.3: Exemple de l'entrée #acm dans le schéma de la base de données*

Une fois que nous avons recueilli les hashtags et les définitions rédigées par les utilisateurs en anglais naturel, nous procédons à la deuxième étape qui consiste à calculer la distance entre les définitions de chaque hashtag.

3.2.2.2 Calcul de distance entre les définitions d'un hashtag

Les définitions rédigées par les utilisateurs pour un hashtag donné peuvent être redondante, c'est-à-dire que certaines définitions peuvent décrire le même sens. Notre objectif dans cette étape est de mesurer la similarité sémantique entre les définitions afin de fournir une entrée pour la phase ultérieure du clustering (c.f. section 3.2.2.3). Pour effectuer le clustering, nous avons besoin d'une métrique définissant la distance entre définitions.

Dans la littérature, les approches traditionnelles pour comparer deux phrases reposent sur la fréquence de co-occurrence des lettres et des termes employés dans les différentes phrases en langage naturel [146, 41]. Ces approches bien connues en recherche d'information [147, 120, 148, 149, 150, 151] sont limitées à la stricte co-occurence des mêmes termes dans les définitions. Cependant les définitions recensées sur les dictionnaires en ligne sont générées par différents utilisateurs de différentes origines, cultures et même avec des langues maternelles différentes. Ceci rend l'ensemble des termes composant les définitions hétérogènes. De plus, ces définitions issues du Web 2.0 incluent un grand nombre de néologismes et des abréviations. Pour surmonter ces limitations, nous avons besoin d'une base de connaissances externe, permettant de prendre en considération la proximité entre les termes dans la métrique de distance entre deux définitions. Ce problème est considéré comme un problème de similarité sémantique pour la désambiguisation du sens d'un hashtag [152].

Parmi les différentes techniques faisant appel à une base de connaissances externe, l'algorithme Extended Lesk s'est avéré être l'un des plus efficaces [153]. Cet algorithme est une adaptation de l'algorithme Lesk [154] utilisant Wordnet[3] comme base de connaissances externe. Extended Lesk améliore significativement le résultat de calcul de similarité vu qu'il prend en compte les liens et relations taxonomiques WordNet dans la définition d'un sens donné, et le calcul de similarité entre les définitions se basent sur les différents contextes des mots qui les composent.

Extended Lesk fournit le sens (synset) le plus plausible pour un mot parmi tous les sens possibles de ce mot dans Wordnet en utilisant les l'ensemble des mots qui l'entoure dans la phrase. Pour cela, l'algorithme calcule la distance entre les candidats possibles (l'ensemble des synsets contenant le terme à désambiguïser), et sélectionne le plus proche en moyenne de tous les termes appartenant à l'ensemble de des mots entourant le mot à désambiguïser, qui sont aussi présents dans Wordnet. Ceci dit qu'il

3. http ://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement

est possible d'identifier le synset associé à un mot dans une définition donnée d'un hashtag à condition que ce terme existe dans Wordnet.

Extended Lesk est limité à la similarité sémantique entre deux mots. Dans [155], les auteurs proposent une nouvelle approche pour la similarité sémantique entre deux phrases en utilisant l'algorithme Extended Lesk. À cette étape de notre approche, nous effectuons un pré-traitement utilisant Extended Lesk pour calculer la distance entre les définitions d'un hashtag, (voir la section 3.2.1). En supposant que A et B sont deux définitions, et que $\#A$ (resp. $\#B$) est le nombre de mots dans chaque phrase A (resp. B), cette approche fonctionne comme suit :

- La première étape consiste à construire une matrice de similarité MS[$\#A$, $\#B$] pour chaque paire de mots à désambiguisés. MatSim[a, b] représente la similarité sémantique entre le sens le plus plausible pour le mot (trouvé à l'aide d'Extended Lesk) à la position a dans A et le mot à la position b dans B.
- La deuxième étape vise à calculer la similarité entre deux définitions d'un hashtag donné. Il s'agit de trouver le mariage entre l'ensemble de mots qui composent la définition A et l'ensemble de mots qui composent la définition B, de manière à ce que la valeur issu de ce mariage soit maximisée [156]. Ce problème est donc ramené à un problème de mariage dans un graphe bipartite pondéré et peut être résolu grâce à la méthode hongroise d'affectation [157] dont la formule est 3.7.

$$\frac{2 \times \sum_{i=1}^{Min|A|\times|B|} Match(A_a, B_b)}{|A|+|B|} \tag{3.6}$$

Où $Match(A_a, B_b)$ correspondant à la valeur de correspondance entre les termes de la $i^{ème}$ paire trouvé dans le mariage maximisant le poids total des assocations entre les mots de la définition A et ceux de la définition B.

Exemple si $|A| = 3$ et $|B| = 5$, et que la méthode hongroise nous génère trois couples de termes avec des similarités respectives de 0.3, 0.6 et 0.9 la similarité sémantique entre les deux définitions A et B sera égale à :

$$\frac{2 \times (0.3+0.6+0.9)}{3+5} = 0.45 \tag{3.7}$$

Ainsi les différentes valeurs de similarités calculées entre les différentes paires de définitions d'un même hashtag construisent la matrice d'adjacence (figure 3.4) d'un graphe pondéré dont les sommets sont les définitions d'un hashtag, et les arêtes sont

pondérés par la distance entre les deux définitions.

		Hashtag1			
		definition1	définition2	définition3	définition4
hashtag1	definition1	1	distance	..	..
	définition2	..	1	..	..
	définition3	..	..	1	..
	définition4	..	..	..	1

FIGURE *3.4: Matrice de Distance entre les définitions d'un hashtag*

3.2.2.3 Clustering de définitions

L'objectif de cette étape et d'alimenter $S(w)$, $\forall w \in W$. Une matrice **simMeaning** pour chaque hashtag est ensuite utilisé pour regrouper les définitions d'un hashtag en fonction de leurs significations. En d'autres termes Lors de cette étape, nous construisons notre folksionary qui regroupe en unités de sens les définitions des hashtags recensées sur les différents dictionnaires en ligne, en unité de sens.

Dans notre approche nous n'avons pas d'information a priori concernant le nombre de clusters à générer. Cependant le nombre de clusters à générer est une information nécessaire pour les algorithmes de clustering standard. Pour palier à cette limitation nous avons opté pour un algorithme de clustering à base de graphe. Nous avons choisi de comparer les cinq algorithmes de clustering à base de graphe proposés dans [158] : à savoir Markov Clustering (**MCL**), Iterative Conductance Cutting, Geometric (**ICC**), The Geometric MST Clustering algorithm (**GMC**), The Normalized Cut Clustering algorithm (**NCC**), et une extension du The Fuzzy C-Means MST Clustering algorithm (**FMC**) développée par les auteurs du [158] (voir tableau 3.1).

A part l'algorithme FMC où les auteurs ont effectué des réglages pour générer automatiquement la combinaison optimale des paramètres. tous les autres algorithmes tous les autres algorithmes nécessitent des paramètres à fournir à l'avance. En ce qui concerne la mémoire, tous les algorithmes à l'exception de GMC et FMC adapté sont exprimées en utilisant des matrices dérivés de la matrice d'adjacence du graphe, ils dominent donc des besoins en mémoire qui impactent significativement le temps

d'exécution. Tandis que les algorithmes basés sur le Spanning Tree minimum (FMC et GMC) le coût en terme de mémoire et de temps sont beaucoup plus réduits.

	Caractéristiques			
Algorithme	Complexité(T)	Complexité(E)	Critère d'arrêt	Paramètres utilisateur
FMC	O(E$\log$N)	O(N+E)	$\mathcal{E}$ =0.5	totalment non supervisé
MCL	O(N^3)	O(N^2)	convergence aléatoire	r=2, k=3
ICC	O(N^3)	O(N^2)	conductance > α=0.45	α
GMC	O(N^3)	O(N+E)	MST > λ	λ
NCC	O(N^2)	O(N^3)	coupure normalisée > α	$\alpha = 0.7$, $\sigma = 0.4e_{max}$

Tableau 3.1: Etude comparative des algorthmes de clustering à base de graphe

Pour conclure, nous avons opté dans notre approche pour l'algorithme de Markov Clustering (MCL) [159] du fait qu'il demeure une technique de clustreing remarquablement robuste aux techniques classiques. Il n'exige pas de de spécifier le nombre de clusters à trouver par avance, grâce à l'alternance des deux techniques de l'expansion et l'inflation sur valeurs dans la matrice représentant les éléments à clusteriser. Et également parce qu'il n'exige pas de critères d'arrêt comme la convergence se fait de manière aléatoire. Cependant le grand challenge de MCL reste le choix des valeurs des paramètres en entrée par l'utilisateur à savoir :

- **gammaExp** le paramètre d'inflation, qui influe sur la granularité des clusters,
- **maxResidual** la mesure de l'idempotence pour la terminaison de l'algorithme,
- **maxZero** la valeur maximale en dessous de laquelle une valeur sera considérée comme nulle dans les opérations d'élagage.

Nous appliquons l'algorithme Markov Clustering pour regrouper les définitions d'un hashtag en des clusters que nous interprétons comme étant des unités de sens. Ainsi grâce à sa technique de marches aléatoires dans un graphe, MCL permet d'identifier les zones d'un graphe de similarité les plus fortement connectées. Notre algorithme de clustering peut ainsi s'écrire comme dans l'algorithme [1].

Données: crawl : collection(String,collection(String)) // ensemble de définitions pour tous les hashtags

Résultat: Folksionary :collection(String,collection(String,String)) // ensemble des unités de sens pour tous les hashtags

pour *chaque hashtag w dans crawl* **faire**
- k : float ; k⟵0
- distance : float[][] ; Defs : collection(String)
- SimilarityMatrix : (String,float[][],collection(String))// structure contenant le hashtag w, la matrice de ditance entre ses définitions et la liste de ces définitions
- FinalClusters : collection(collection(Integer)) // ensemble des clusters générés par Markov Clsutering
- cluster : collection(Integer) ; defId : Integer ; def : String
- numCluster : int ; numCluster ⟵ 1 sensMap : collection(Integer,String) ; / ensemble de sens du hashtag w
- distance=AdaptativeLeskAlgo.calculateDistance(w) ; // génère la matrice de distance entre les définitions du hashtag w
- Defs⟵ getDefinition(w) ; // renvoie les définitions du hashtag w
- SimilarityMatrix.put(w,distance,Defs) ;
- FinalClusters ⟵mcl(SimilarityMatrix,gammaExp,maxResidual,maxZero) ;
- **pour** *chaque cluster dans FinalClusters* **faire**
  - **pour** *chaque defId dans cluster* **faire**
    - def ⟵ similarityMatrix.getDefinitions().getValue(defId) ;
    - sensMap.addSense(numCluster, def) ;
    - MeaningList.put(sens+(k+1),SensMap) ;
  - **fin**
  - numCluster ⟵ numCluster+1
- **fin**
- **si** *sensMap.getsenses.isempty()* **alors**
  - **pour** *chaque def dans similarityMatrix.getDefinitions()* **faire**
    - sensMap.addSense(1, def) ;
  - **fin**
- **fin**
- Folksionary.put(w,sensMap) ;

fin

return Folksionary ;

Algorithm 1: Algorithme de génération de Clusters de définitions (Folksionary)

Lors de cette étape nous réalisons la fonction de clustering noté précédemment $\mathcal{E}$ pour générer l'ensemble de sens noté $S(w)$. Nous sommes alors en mesure de percevoir dans quelle mesure chaque hashtag est polysémique et avec quel cardinal(figure 3.5).

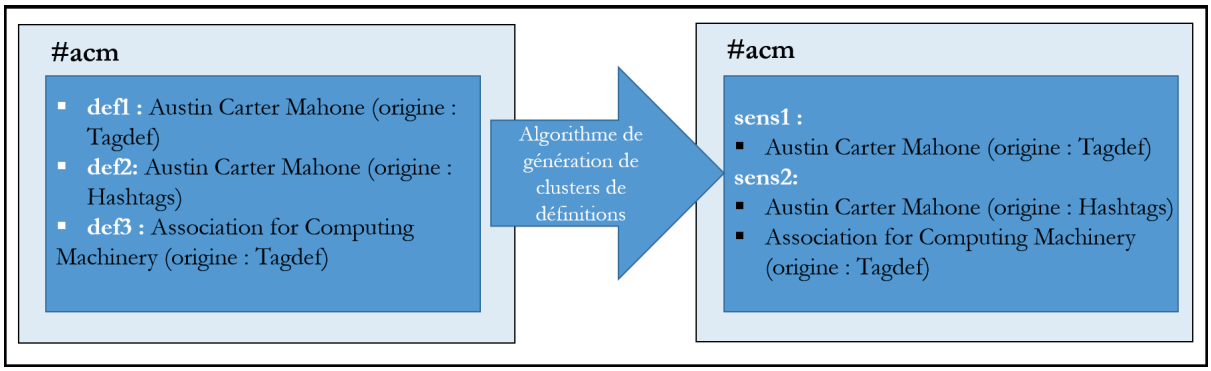

FIGURE *3.5: Clustering des définitions de l'entrée #acm dans le folksionary*

Dans l'étape suivante nous générons une sortie lisible par l'humain, avec un apparence semblable à un dictionnaire traditionnel.

3.2.2.4 Clustering hiérarchique des sens des hashtags

Dans cette section nous procédons à une première instance de clustering hiérarchique [160, 161] appliqué au Folksionay. Comme déjà expliqué précédemment dans la section formalisation, l'une des perspectives de notre travail est de découvrir les relations qui puissent exister entre les hashtags pour deux principales raisons : (1) générer une ontologie à partir du folksionary et (2) étendre les requêtes au niveau du folksionary en se basant sur les relations de similatité entre hashtags (synonymies), et/ou spécialiser (resp. généraliser) les requêtes grâce au relations de méronymies (resp. hypernymies).Dans ce travail nous nous focalisons à la découverte de relations de synonymie. Le clustering hiérarchique vise à créer une décomposition hiérarchique des objets selon certains critères. Il nécessite deux éléments essentiels : une mesure de distance et une approche de fusion de données. Le principe du clustering hiérarchique est de chercher les deux points les plus proches selon la distance et on les regroupe dans un cluster, les points sont remplacés par leur centre. Ensuite on continue à chercher à nouveau les points (ou clusters) les plus proches pour les regrouper

en un cluster, et ceci de manière itérative. Pour chaque calcul de distance entre deux classes, la méthode de fusion est utilisée (voir méthodes de fusion ci-après). Et on continue à itérer jusqu'à n'avoir qu'une seule classe.

Les méthodes de fusion de clusters possibles :

- **lien simple** : on prend le minimum entre deux points de chaque classe (simple, mais à tendance à tout agglomérer de proche en proche).
- **lien complet** : on prend le maximum entre deux points de chaque classe.
- **moyenne** : on prend la moyenne de la distance entre toutes les paires.
- **ward** : permet de trouver à chaque étape l'agrégation qui minimise la perte d'information (produit des classe plus compacts).

Notre objectif est de faire une première classification hiérarchique des hashtags afin de découvrir une simple relation de synonymie. Nous avons opté pour une méthode simple et largement utilisé, la classification hiérarchique ascendante. A ce stade nous procédons au clustering hiérarchique entre les unités de sens d'un hashtags, sachant qu'une unité de sens est un ensemble de définitions partageant le même sens. Nous optons pour la même mesure de distance utilisée précédemment (c.f. section 3.2.2.2) étant donné qu'on a déjà vérifié son efficacité sur nos données. Pour la méthode de fusion nous utilisons la moyenne des distances étant donné sa simplicité d'implémentation et du fait qu'à ce stade l'objectif est de vérifier l'existence de relations entre les entrés du folksionary. Dans la suite de nos travaux nous opterons pour un apprentissage automatique afin de choisir la meilleure méthode de fusion adaptée à nos données.

Dans la section précédente nous avons calculé la distance entre les définitions d'un même hashtag (c.f. section 3.2.2.2), ensuite nous avons effectué le clustering (c.f. section 3.2.2.3) afin de pourvoir les regrouper en unités de sens (voir figure 3.5). Ici, l'objectif est de faire un clustering de sens inter hashtags. Nous commençons par le calcul de distance entre les sens des hashtags deux à deux. Prenons l'exemple de deux hashtags : **#100factsaboutme** ayant trois sens et le **#10212011** avec deux sens (figure 3.6). Le calcul de distance entre deux sens donnés revient à calculer les distance entre les définitions de ces deux sens (figure 3.8). Cette matrice est ensuite transformée en moyenne pour exprimer la distance entre deux sens distincts(figure 3.7). Nous utilisons la notation nom du hashtag concaténé avec un chiffre pour représenter le sens d'un hashtag e.g #10212011==>1 désigne le 1^{er} sens du hashtag #10212011.

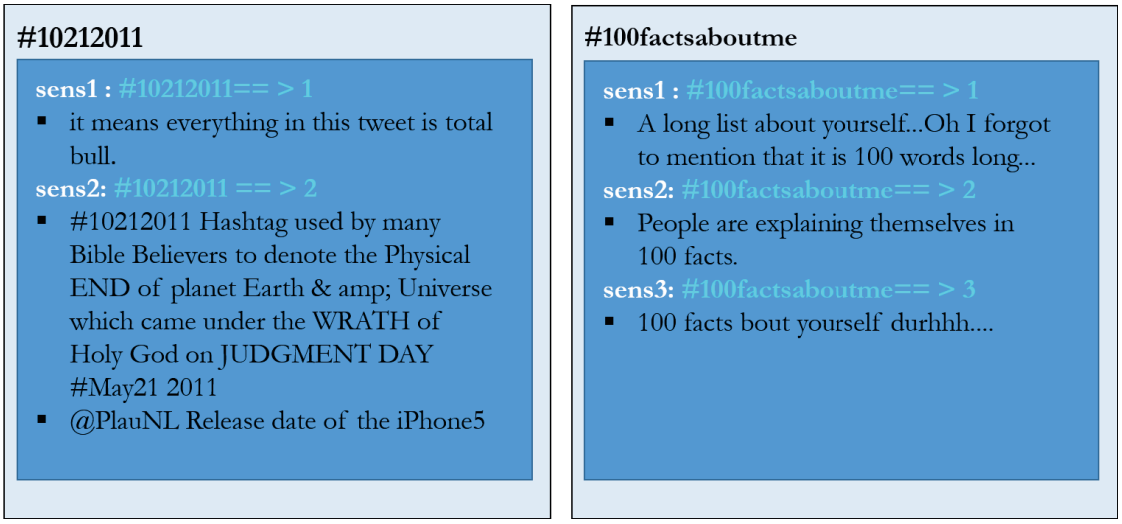

FIGURE *3.6: les entrées #10212011 et #100factsaboutme dans le folksionary*

#10212011 (hashtag2)	#100factsaboutme (hashtag1)		
	#100factsaboutme== > 1 (sens1)	#100factsaboutme== > 2 (sens2)	#100factsaboutme== > 3 (sens3)
#10212011== > 1 (sens1)	distance	....	....
#10212011== > 2 (sens2)	....	....	distance

FIGURE *3.7: Matrice de Distance entre les sens de #10212011 et #100factsaboutme*

#10212011 == >2 (sens2)	#100factsaboutme == > 1 (sens1)
	définition 1 : A long list about yourself...Oh I forgot to mention that it is 100 words long...
définition 1 : #10212011 Hashtag used by many Bible Believers to denote the Physical END of planet Earth & amp; Universe which came under the WRATH of Holy God on JUDGMENT DAY #May21 2011	distance
définition 2 : @PlauNL Release date of the iPhone5	....

FIGURE *3.8: Matrice de Distance entre les sens #10212011==>2 et #100factsaboutme==>3*

Nous passons ensuite au clustering hiérarchique ascendant (HAC)(voir algorithme 2). Le HAC fonctionne tel que suit : chaque classe (cluster) est progressivement "absorbée" par la classe la plus proche jusqu'à la fusion des deux derniers clusters.

Données:
Sens$_{Set}$: collection(collection(Integer,String)) // `ensemble de tous les sens de tous les hashtags du folksionary`

Résultat:
Cluster : collection(Integer,String) // `chaque sens est placé dans son propre cluster`
Clusters : collection(String,Sens$_{Set}$) // `ensemble initiale de clusters`

ClustDistances : collection(double, collection([] String)) // `matrice de distance de tous les sens du folksionary`
clusterPair : collection([] String) // `la paire de sens pour lesquels nous calculons la similarité`
ClustersFinal : collection(String, clusterPair)// `ensemble de clusters générés par le Clustering Hiérarchique`
pour *chaque sens* $\in$ *Sens*$_{Set}$ **faire**
  Cluster $\leftarrow$ sens
fin
i, j : float ; distance : float ; clusterName : [] String ; i $\leftarrow$ 0 ; j $\leftarrow$ i+1 : clusterName $\leftarrow$ ""
pour *chaque Cluster*$_i$ $\in$ *Clusters* **faire**
  pour *chaque Cluster*$_j$ $\in$ *Clusters* **faire**
    distance $\leftarrow$ computeSimilarityBetweenClusters(Cluster$_i$,Cluster$_j$)) ;
  fin
  clusterPair.add(Cluster$_i$) ; clusterPair.add(Cluster$_j$) ;
  ClustDistances.put(distance,clusterPair) ;
fin
pour *chaque clusterPair* $\in$ *ClustDistances* **faire**
  si *mostSimilarClusters(clusterPair)* **alors**
    mergeCluster(clusterPair.get(1),clusterPair.get(2)) // `merge les deux clusters dont la similarité est maximale dans un même cluster`
  fin
  clusterName $\leftarrow$ clusterPair.get(1)contact(clusterPair.get(2)) ;
  ClustersFinal.put(clusterName,clusterPair)
fin

Algorithm 2: Algorithme de Clustering Hiérarchique appliqué au folksionary

La figure 3.9 représente une sortie de l'algorithme du clustering hiérarchique appliqué à notre folksionary. La hauteur d'un cluster dans le dendrogramme est égale à la similarité entre les deux clusters avant fusion, tandis que les feuilles représentent le sens des hashtags. Nous avons attribué à chaque sens un nom composé du hashtag auquel il appartient et son numéro dans l'ensemble des sens du hashtag. L'échelle à droite désigne l'indice de hiérarchie. Dans l'exemple de sortie ci-après nous avons trois hashtags ***100wordchallenge***, ***100factsaboutme*** et ***10212011***.

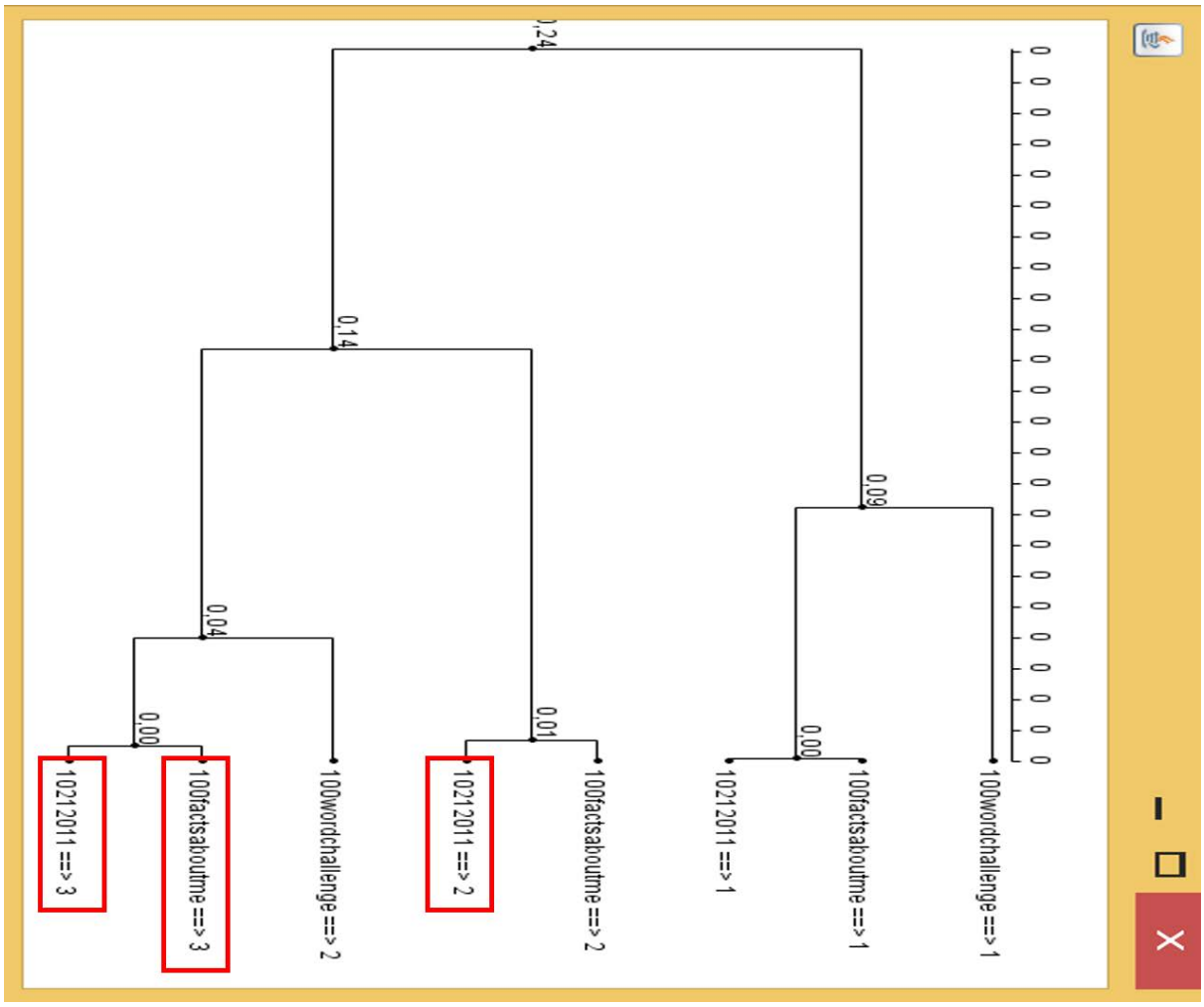

FIGURE *3.9: Résultat du clustering hiérarchique appliqué au folksionary*

3.2.2.5 Formatage du folksionary

L'un des objectifs du folksionary est de fournir un dictionnaire pour des utilisateurs humains. Par conséquent, la sortie est structurée sous un format proche d'un dictionnaire traditionnel. Cette sortie est un fichier qui organise les entrées des hashtags dans un ordre alphabétique. Chaque hashtag est présenté avec tous ses sens, et nous listons pour chaque sens l'ensemble des définitions qui le constituent.
Par exemple, considérons les définitions suivantes, recensées pour le hashtag #acm à l'étape 1(cf. 3.2.2.1)

- "Austin Carter Mahone"
- "Association for Computing Machinery"
- "Austin Mahone :)"

Nous présentons la sortie sous le format d'un dictionnaire classique standard :

> **#acm** **-1.** Association for Computing Machinery **-2.** Austin Carter Mahone ❖ Austin Mahone :)

Comme illustré dans l'exemple précédent, deux sens ont été trouvés pour le hashtag #acm. Le 1^{er} pour acm avec une définition et le 2^{eme} pour le personnage Austin Carter avec deux définitions. Les différents symboles sont définis comme suit :

- les différents sens $s \in S(w)$ sont séparés par des numéros. "**-1.** " indique le début du premier sens. "**-2.** " indique le début du deuxième sens, et ainsi de suite.
- Les définitions partageant le même sens sont séparés par le symbole ❖ .

Le folksionary sous format PDF contenant tous les hashtags avec leurs définitions est disponible en ligne à l'adresse : *http://datasets-satin.telecom-st-etienne.fr/mghenname/folksionary/folksionary.pdf*. Ce folksionary contient 22738 hashtags, et un total de 28 191 définitions.

La figure 3.10 donne un extrait du folksionary à la lettre "G" pour les hashtags de #g2 à #gadgetvader. Par exemple le hashtag#ga possède quatre sens avec une définition par sens, le hashtags #gabos possède deux sens avec deux définitions dans le 1^{er} sens et une définition dans le 2^{eme} sens.

La section suivante est consacré à la présentation des caractéristiques du folksionary généré.

G

#g2 **-1.** Guy that get money and alot of women **-2.** G2 - Gay Men's Mobile Communication

#ga **-1.** General Aviation, all aviation and flying that isn't the airlines or military. **-2.** Grammy Award **-3.** grey's anatomy **-4.** #GA means "good afternoon"

#gabos **-1.** go and bum oso special ❖ Go and bone on saturday **-2.** Game Aint Based On Sympathy

#gabs **-1.** Great American Bake Sale: A national campaign that mobilizes Americans to end childhood hunger by holding bake sales in their communities. ❖ Go And Break Stuff

#gac **-1.** Global Awareness Consulting **-2.** Ghost Adventures Crew

#gac2011 **-1.** gooey excrements that have been forgotten through the years. for example " i played that girl again like some gac2011" **-2.** Ghost Adventures Crew 2011

#gadgetshow **-1.** British tv show all about the latest gadgets and tests them. **-2.** TV show apparently about any thing other than gadgets. **-3.** A justifiable and legal way to leer at Suzi Perry every Monday

#gadgetvader **-1.** jason bradbury off of the @thegadgetshow invented this hash to promote his marble arch annimation internet viral. **-2.** Hashtags or #tags basically a quick simple way to make your twitter posts searchable by anybody and the word that follows the # is what is used for the search. So #gadgetvader is searchable but # gadgetvader is not because of the space between the # and the tag. Written by @1hale

FIGURE *3.10: Sortie Standard du folksionary à la lettre G*

3.2.2.6 Caractérisation du folksionary

Dans cette section nous présentons quelques statistiques du folksionary :

- Le folksionary présente un pourcentage de $\sim 7.94\%$ d'un dictionnaire d'anglais *Oxford Dictionary* a 355 000 entrées.
- Notre approche a identifié 25,106 significations sur 28191 définitions.
- Chaque hashtag a une moyenne de $\sim 1,1$ sens avec un écart type de $\sim 0,45$.
- Dans ce folksionary, 1731 hashtags sur 22 738 ont plusieurs significations.

La figure 3.11 ci-après illustre la répartition des lettres initiales des hashtags en comparaison avec le dictionnaire d'anglais *Oxford Dictionary* tel que rapporté par.[4]

Concentrons-nous sur les 1731 hashtags qui sont polysémiques (ayant différents sens). Parmi les hashtags polysémiques nous avons une moyenne de ~ 2.37 sens par hashtag avec une déviation standard de $\sim 0,94$. La figure 3.12 représente le nombre de hashtags regroupées par nombre de sens.

4. http://en.wikipedia.org/wiki/Letter...

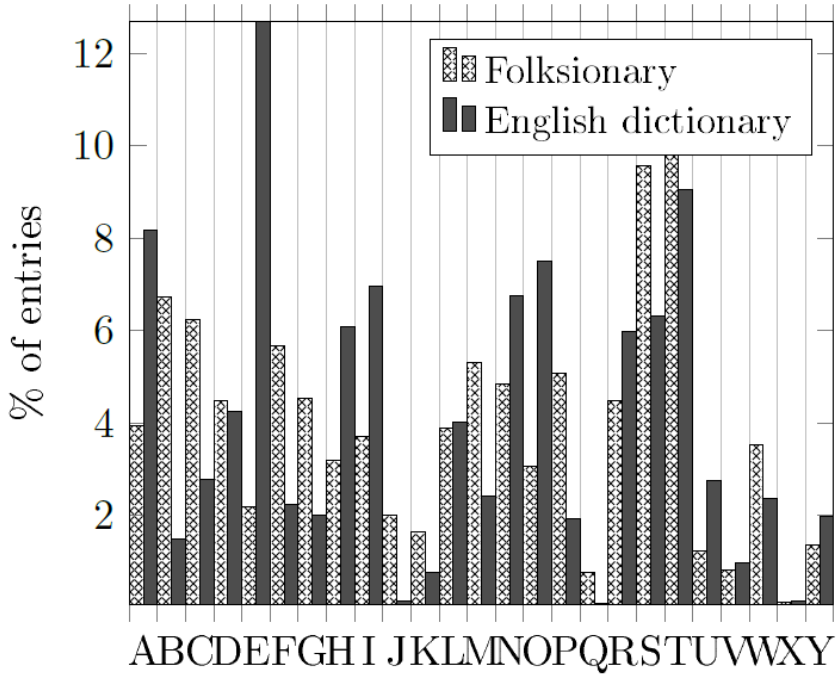

FIGURE *3.11: Les hashtags dans le folksionary en comparaison avec Oxford.*

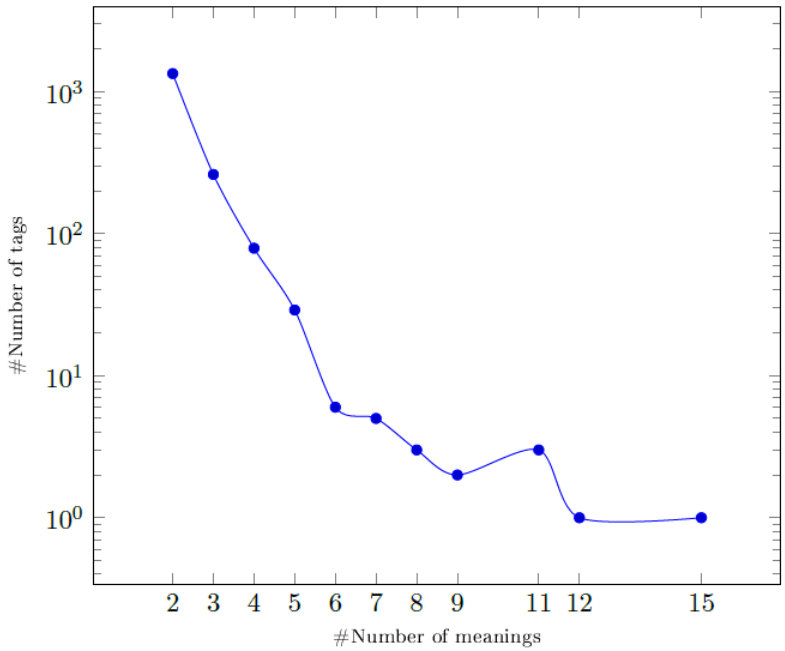

FIGURE *3.12: Nombre de hashtags regroupées par nombre de sens.*

Par exemple : 261 hashtags sont polysémiques avec trois sens distincts détectés par notre approche. Les derniers 1,2% hashtags polysémiques restant avec un degré supérieur ou égal à 6, représente une petite partie de notre folksionary et sont considérés comme des exceptions dans ce travail. Ces hashtags sont généralement les hashtags les plus populaires, tels que *#justinbieber* où les gens expriment des opinions différents souvent ironiques.

Notre approche est sensible aux opinions : les hashtags populaires dans le folksionary doivent avoir des sens différents en termes de concepts auxquels les sens se réfèrent. Ceci est particulièrement vrai pour les hashtags avec beaucoup de sens. Notre approche classe des opinions différentes dans des sens différents.

Dans cette partie, nous avons présenté une analyse quantitative du folskionary généré par notre approche automatique. Dans la section suivante nous traitons l'analyse qualitative du folksionary.

3.3 Prototype et évaluation du folksionary

Cette section est dédiée à notre prototype et à l'analyse qualitative mesurant la distance entre le folksionary généré par notre approche automatique et la vérité du terrain établie manuellement.

3.3.1 Prototype et implémentation

Nous avons construit un ensemble de données via l'extraction de données à partir des sources web. Pour ce faire, nous avons créé un Web scrappers à l'aide de la bibliothèque `pjscrape` javascript [5].Nous avons utilisé Apache Tika [162] pour le filtrage de la langue afin de ne sélectionner que les définitions en anglais. Ensuite, nous calculons la distance à l'aide de notre propre implémentation en Java d'Extended Lesk. Et finalement nous avons réalisé le partitionnement automatique des définitions avec l'algorithme Markov Clustering en utilisant JavaML [163]. Dans la section suivante nous expliquons la démarche suivie pour évaluer le folksionary et expliquer les résultats obtenus.

5. http://nrabinowitz.github.com/pjscrape/

3.3.2 Évaluation

Dans la continuité de notre travail nous avons procédé à la confirmation de la pertinence des résultats obtenus par notre approche de construction de folksionary via une évaluation. Le problème posé était le suivant : ***Comment mesurer l'efficacité du clustering de définitions ?***

De plus nous n'avons pas trouvé non plus de dataset qui nous aurait permis de confronter directement nos résultats avec la littérature. Ceci nous a amenés à deux tâches : la définition de la méthode d'évaluation et la construction d'une vérité de terrain représentant la référence à atteindre.

3.3.2.1 Établissement de la vérité du terrain

Nous avons mis en place une application Web en ligne où les participants peuvent regrouper manuellement les définitions des hashtags en unités de sens. La figure 3.13 présente l'interface utilisateur de cette application. L'évaluation de la pertinence de la méthode utilisée doit être faite en comparant le résultat produit avec des données de références. Nous n'avons pas trouvé dans la littérature de Framework d'évaluation de clustring qui correspond à notre cas lorsque le nombre de clusters n'est pas connu à l'avance. Cette application présente à l'utilisateur le hashtasg et l'ensemble de ses définitions. L'utilisateur peut :

- créer un sens ce qui rajoute une colonne Meaning(cela correspond à créer un cluster)
- placer une définition dans un sens, ce qui peuple le cluster correspondant.

Le travail est répété pour l'ensemble des hashtags du dataset.

Dans l'exemple donné figure 3.13 l'utilisateur a défini trois sens pour les définitions du hashtag #1omf, le premier(Meaning 1) contient deux définitions (#1 Of My Followers, 1 Of My Followers), le second (Meaning 2) contient une définition (1 of my freinds), et le troisième une définition également (i have no idea).

Ce travail correspond à l'établissent manuel d'un partitionnement des définitions c'est donc une fonction de clustering $\mathcal{E}$ (cf.3.1) comme présentée dans la section (cf.3.2.1). Nous la notons $\mathcal{E}_{GT}$ pour partitionnement Ground Truth ou vérité du terrain. Nous avons fait intervenir dix personnes dans la construction du Ground Truth afin d'éviter au maximum la subjectivité et créer un partitionnement manuelle le plus précis possible.

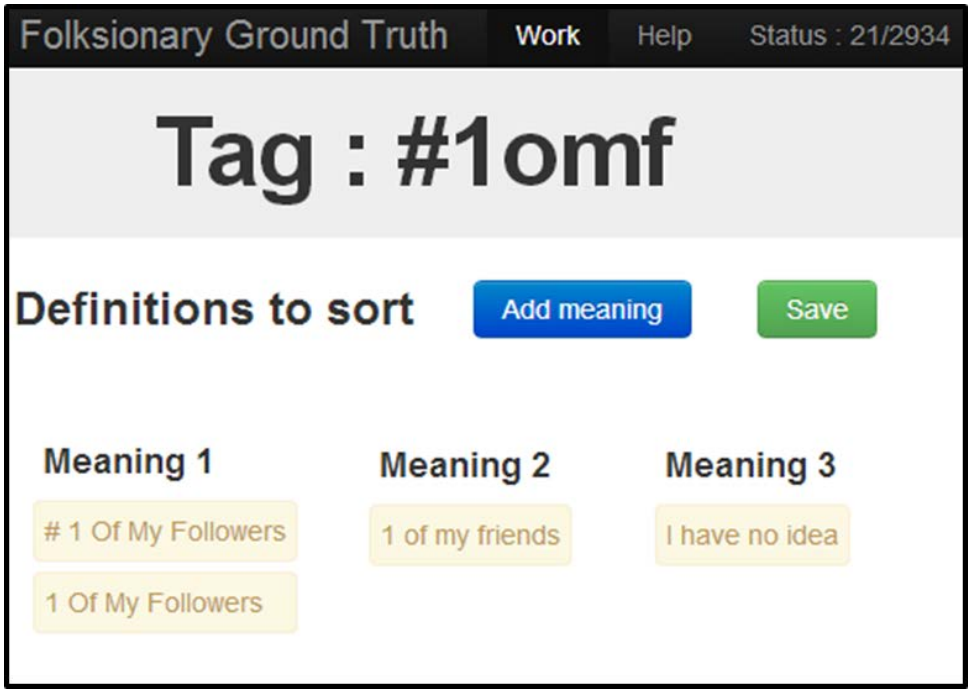

FIGURE *3.13: Application Web pour la construction de la vérité de terrain*

3.3.2.2 Protocole d'évaluation par paires

La validation du clustering étudie pour chaque hashtag, le clustreing des définitions pour chaque paire de défintions d'un hahstag. Nous utilisons les notations suivantes :

- $\mathcal{E}_{GT}$ le partitionnement construit manuellement.
- $\mathcal{E}_P$ le partitionnement prédit par notre approche.

L'objectif de l'évaluation est de mesurer la distance entre $\mathcal{E}_P$ et $\mathcal{E}_{GT}$. Nous utilisons une évaluation par paire.

Définition 9

$S_{GTi} = \mathcal{E}_{GT}(d_i)$ *(respectivement.* $S_{GTj} = \mathcal{E}_{GT}(d_j)$*).*

$S_{Pi} = \mathcal{E}_P(d_i)$ *(respectivement.* $S_{Pj} = \mathcal{E}_P(d_j)$ *).*

le clustering est bon si :

- les deux définitions sont dans le même cluster dans le $\mathcal{E}_P$ et aussi dans le $\mathcal{E}_{GT}$
- les deux définitions sont dans des clusters différents dans le $\mathcal{E}_P$ et aussi dans le $\mathcal{E}_{GT}$

Le clustering est mauvais si :

– les deux définitions sont dans le même cluster dans l'un des partitionnements mais séparées dans l'autre

Ceci nous conduit à définir les observations suivantes :

– Vrai positif (TP) si : $S_{GTi} = S_{GTj}$ et $S_{Pi} = S_{Pj}$;
– Vrai négatif (TN) si : $S_{GTi} \# S_{GTj}$ et $S_{Pi} \# S_{Pj}$;
– Faux positif (FP) si : $S_{GTi} \# S_{GTj}$ et $S_{Pi} = S_{Pj}$;
– Faux négatif (FN) si : $S_{GTi} = S_{GTj}$ et $S_{Pi} \# S_{Pj}$.

Les mesures d'évaluation les plus classiques qu'on peut trouver dans la littérature sont le $F_1 score$ (cf. équation 3.8) et le coefficient de corrélation Matthews (MCC) (cf. équation 3.10). Mais le problème avec ces deux mesures est qu'elles sont indéfinies lorsque la valeur du dénominateur est nulle, ce qui arrive assez souvent dans notre cas.

Définition 10

$$F_1 score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.8}$$

tel que :

$$Precision = \frac{|TP|}{|TP| + |FP|} \quad et \quad Recall = \frac{|TP|}{|TP| + |FN|} \tag{3.9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.10}$$

De ce fait nous avons opté pour l'Average Conditional Probability (ACP) [164] (cf. équation 3.11) qui permet de prendre en compte un tel cas. ACP est défini comme suit si toutes les sommes sont non nulles :

Définition 11

$$\begin{aligned} ACP = & \frac{1}{4}[\frac{|TP|}{|TP| + |FN|} + \frac{|TP|}{|TP| + |FP|} + \frac{|TN|}{|TN| + |FP|} \\ & + \frac{|TN|}{|TN| + |FN|}] \end{aligned} \tag{3.11}$$

Sinon, ACP est la moyenne des probabilités conditionnelles qui sont définies.

L'algorithme 3 donne les étapes pour l'évaluation du clustering.

Data:
$crawl$: collection(String,collection(String)) // hashtags et définitions dans la BD

Result:
Effectiveness : collection(String,Float) // le rapprochement entre GT et DP
GT : collection(String,collection(String)) // GT the Dataset Ground Truth partitioning

DP : collection(String,collection(String)) // DP the Dataset Prediction partitioning

pour *chaque hashtag* $w \in GT$*.keys()* **faire**
| $GT(w)$=$\mathcal{E}_{GT}(crawl(w))$
fin
pour *chaque hashtag* $w \in DP$*.keys()* **faire**
| $DP(w)$=$\mathcal{E}_{DP}(crawl(w))$
fin
$TP \leftarrow 0$; $TN \leftarrow 0$; $FP \leftarrow 0$; $FN \leftarrow 0$
pour *chaque hashtag* $w \in crawl.keys()$ **faire**
| $pairDefs : collection(string)$
| $pairDefs \leftarrow getPairwiseDefCombinaison(w)$
| **pour** *chaque paire* $(d_i, d_j) \in pairDefs$ **faire**
| | $C(s_{GTi}) \leftarrow$ retrieveClusterSense($\mathcal{E}_{GT}, d_i$) ;
| | $C(s_{GTj}) \leftarrow$ retrieveClusterSense($\mathcal{E}_{GT}, d_j$)
| | $C(s_{Pi}) \leftarrow$ retrieveClusterSense($\mathcal{E}_{P}, d_i$)
| | $C(s_{Pj}) \leftarrow$ retrieveClusterSense($\mathcal{E}_{P}, d_j$)
| | **si** $C(s_{GTi}) = C(s_{GTj})$ *and* $C(s_{Pi}) = C(s_{Pj})$ **alors**
| | | $TP \leftarrow TP + 1$
| | **fin**
| | **si** $C(s_{GTi}) \neq C(s_{GTj})$ *and* $C(s_{Pi}) \neq C(s_{Pj})$ **alors**
| | | $TN \leftarrow TN + 1$
| | **fin**
| | **si** $C(s_{GTi}) \neq C(s_{GTj})$ *and* $C(s_{Pi}) = C(s_{Pj})$ **alors**
| | | $FP \leftarrow FP + 1$
| | **fin**
| | **si** $C(s_{GTi}) = C(s_{GTj})$ *and* $C(s_{Pi}) \neq C(s_{Pj})$ **alors**
| | | $FN \leftarrow FN + 1$
| | **fin**
| **fin**
| $float\ \ ACP \leftarrow calculateACP(TP, TN, FP, FN)$
| Effectiveness.put(w,ACP)
fin
return Effectiveness

Algorithm 3: Algorithme d'évaluation du Clustering

Une vue synoptique de ce processus est tel que suit :

1. pour chaque $w \in W$, énumérer toutes les paires de définitions $(d_i, d_j) \in D(w) \times D(w)$ tel que i < j.
2. Récupérer $S_{GTi} = \mathcal{E}_{GT}(d_i)$ (respectivement. $S_{GTj} = \mathcal{E}_{GT}(d_j)$).
3. $S_{Pi} = \mathcal{E}_P(d_i)$ (respectivement. $S_{Pj} = \mathcal{E}_P(d_j)$).
4. Faire des observations (vrai positif, vrai négatif, faux positif, faux négatif) en fonction des valeurs de $S_{GTi}, S_{Pi}, S_{GTj}, S_{Pj}$
5. Calculer le score global de concordance entre les deux partitionnement pour tous les $w \in W$.

3.3.2.3 Exemple et Interprétation

Notre but est d'étudier le rapprochement entre l'approche automatique du clustering et celle de la verité de terrain établie manuellement. Nous présentons ci-après un exemple de la procédure d'évaluation de la qualité de clustering pour le hashtag #nrg que nous avons choisi au hasard parmi tous ceux qui ayant minimum trois définitions et dont les définitions sont longues et plus au moins compliquées :

Exemple : Soit la situation suivante $\mathcal{S}_1$ où :

$$\mathcal{S}_1 : \begin{cases} \mathcal{W} = \{\#nrg\} \\ D(\#nrg) = \{d_1 = \text{"Energy comes in physical forms such as heat, light, electricity, potential, kinetic or mechanical energy. Energy can be stored in resources such as petroleum,wood or other carbon products, or in uranium. Energy sources that are captured and con-verted/ transmitted as electricity include solar, hydro, tidal and wind."}, \\ \qquad d_2 = \text{"energy=Mental or psychic energy or activity is the concept of a principle of activity powering the operation of the mind or psyche. Many modern psychologists or neuroscientists would equate it with increased metabolism in neurons of the brain."} \\ \qquad d_3 = \text{"NRG is 2 hours of uplifting dance anthems every Friday night with DJ Matt Rogan. More information and live streams at mattrogan.com"}\} \end{cases} \tag{3.12}$$

Le tableau 3.2, représente le partitionnement automatique $\mathcal{E}_P$ et manuel $\mathcal{E}_{GT}$ pour les définitions du hashtag **#nrg** .

$D(w)$	**Partitionnement manuel ($\mathcal{E}_{GT}$)**	**Partitionnement automatique ($\mathcal{E}_P$)**
$\{d_1, d_2, d_3\}$	$\{d_1, d_2\}$ $\{d_3\}$	$\{d_1, d_2\}$ $\{d_3\}$

Tableau 3.2: Partitionnement pour le hashtag #nrg

Le tableau 3.3.2.3 présente la liste exhaustive des observations pour chaque paire.

$\mathcal{E}_{GT}$	$\mathcal{E}_P$	Observation (TP, TN, FP, FN)
$s_{GT_1} = s_{GT_2}$	$s_{P_1} = s_{P_2}$	TP
$s_{GT_1} \neq s_{GT_3}$	$s_{P_1} \neq s_{P_3}$	TN
$s_{GT_2} = s_{GT_3}$	$s_{P_2} = s_{P_3}$	TP

Tableau 3.3: Observations pour l'exemple #nrg

Dans l'exemple #nrg, nous avons les résultats suivants :

- $Recall = 1$
- $Precision = 1$
- $F_1 score = 1$
- $ACP = 1$

3.3.2.4 Évaluation

Nous avons choisi L'algorithme MCL à base de graphes pour notre approche de partitionnement automatique, car il peut être utilisé pour détecter des clusters de différentes formes sans préciser le nombre à l'avance. Mais, en dépit de son efficacité le choix des valeurs des paramètres en entrée représente un véritable défi. Dans cette section, nous présentons des résultats expérimentaux réalisés sur notre folksionary, afin de générer la combinaison de paramètres représentant le meilleur réglage pour

l'algorithme MCL dans notre cas. Après cette mise au point, nous évaluons notre approche de clustering par rapport à la vérité du terrain établie manuellement. L'algorithme MCL a été présenté à la section 3.2.2.3 . Les paramètres de cet algorithme définis précédemment sont : (1)**gammaExp**, (2)**maxResidual**, (3)**maxZero**. MCL utilise des valeurs de paramètres par défaut, qu'il faut modifier et tester pour avoir la meilleure combinaison adaptée à notre jeu de données. Nous effectuons nos expériences pour la gamme des valeurs suivantes :

- maxZero : $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}$.
- maxResidual : $1, 0, 10^{-1}, 10^{-2}, 10^{-3}$.
- gammaExp : 1.4, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20.

Un bon clustering nécessite un bon choix de paramètres. Dans le 1^{er} test reporté sur la figure 3.14 nous avons les es pourcentages des ACP en variant **maxZero** et en fixant **maxResidual** à 10^{-3}. Ainsi en analysant les résultats nous remarquons clairement la valeur d'ACP converge rapidement à de très bonnes valeurs pour les petites valeurs de **gammaExp** et de **maxZero**. Par exemple, pour **maxZero**=10^{-1} la valeur de ACP reste constante à **89,2 %** à partir de **gammaExp**=6 et commence à augmenter progressivement pour atteindre **89,2 %** à partir de 8, 10, 14, 18, 20 pour les valeurs de **maxZero** $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}$ respectivement.

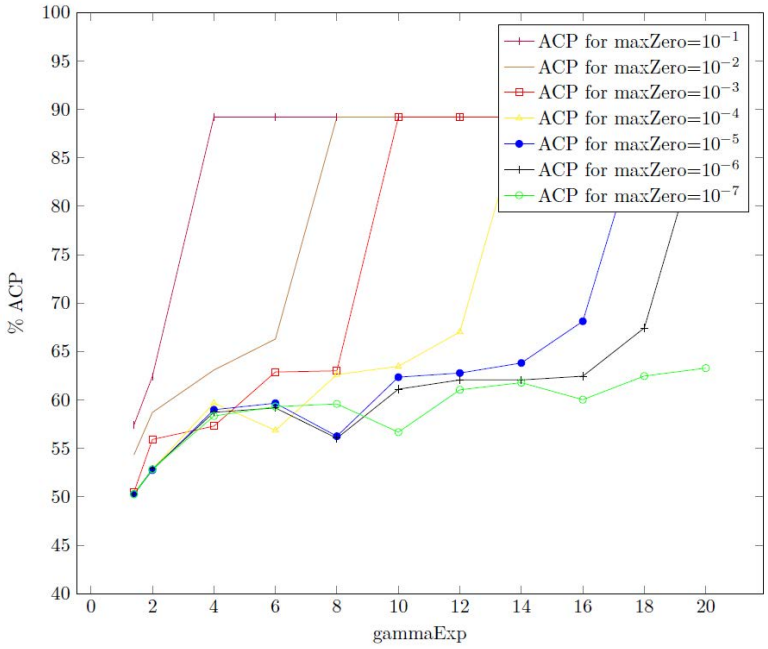

FIGURE *3.14: Pourcentages d'ACP en variant maxZero et fixant maxResidual*

Pour le test qui suit représenté dans la figure La figure 3.15, la valeur **maxZero** est fixée à 10^{-1} tandis que les valeurs de **gammaExp** et **maxResidual** varient.

	maxZero=10^{-1}				
gammaExp	**r=1**	**r=0**	**r=**10^{-1}	**r=**10^{-2}	**r=**10^{-3}
20	53.20	50.98	89.21	89.21	89.21
18	53.20	50.24	89.21	89.21	89.21
16	53.19	50.80	89.21	89.21	89.21
14	53.20	50.09	89.21	89.21	89.21
12	53.20	51.51	89.21	89.21	89.21
10	53.20	49.43	89.21	89.21	89.21
8	53.19	48.05	89.21	89.21	89.21
6	53.20	51.20	89.21	89.21	89.21
4	53.20	60.47	63.78	62.30	62.44
2	52.95	49.94	89.21	89.21	89.21
1.4	53.20	54.79	57.07	57.18	57.41

*Tableau 3.4: L'analyse ACP pour la valeur de **maxZero=0,1***

Comme indiqué dans le tableau précédent la valeur d'ACP reste constante à **53,2%** pour **maxResidual=1** et ne dépasse pas **54,7%** pour **maxResidual=0** quelles que soient les valeurs de **gammaExp** testées. Nous sommes donc focalisés sur le test de **maxResidual** dans l'intervalle [10^{-1}, 10^{-3}].

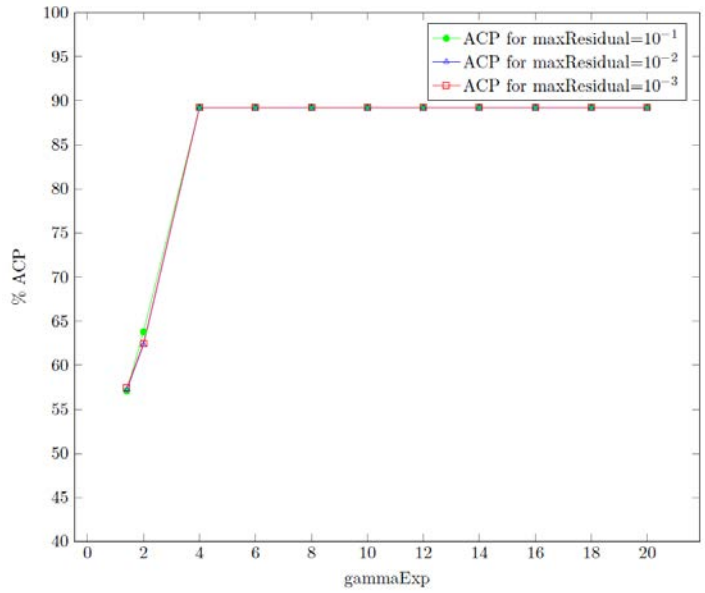

FIGURE *3.15: Pourcentages d'ACP en variant maxResidual et fixant maxZero to* 10^{-1}

Les résultats illustrés dans la figure 3.15 montre que la valeur d'ACP reste constante à **89,2%** à partir de **gammaExp=6** et dans l'intervalle [10^{-1}, 10^{-3}] de **maxResidual**. Pour choisir la meilleure valeur pour les deux paramètres, nous nous sommes appuyés sur un autre critère : la complexité temporelle. Nous avons calculé le temps d'exécution pour différentes valeurs de **maxResidual** avec différentes valeurs de **gammaExp** et nous avons opté pour la combinaison qui converge plus rapidement que les autres. Le tableau 3.5 résume les combinaisons pour les différents tests et leur temps d'exécution.

	maxZero=10^{-1}		
gammaExp	**r=**10^{-1}	**r=**10^{-2}	**r=**10^{-3}
6	34 min 7 sec	32 min 6 sec	31 min 40 sec
20	33 min 0 sec	31 min 23 sec	30 min 53 sec

Tableau 3.5: Temps d'exécution pour les différentes combinaisons de gammaExp et maxResidual

Nous avons poussé la valeur de **gammaExp** à 200 et 2000 et nous avons constaté que plus sa valeur augmente, plus le temps d'exécution diminue. Alors la meilleure configuration de l'algorithme MCL pour nous est : **maxZero=**10^{-1}, **maxResidual =** 10^{-3} et **gammaExp= 20**.

L'analyse expérimentale effectuée confirme que les résultats générés par le partitionnement automatique avec le meilleur réglage de l'algorithme MCL approchent celles provenant de la vérité du terrain avec un pourcentage de 89,2%, ce qui prouve que notre approche pour la classification des définition en unités de sens fournit de bons résultats.

Il convient de noter également que l'évaluation de la performance de notre approche de regroupement n'était pas anodine. La construction manuelle de la vérité du terrain n'est pas une tâche facile. Il y a toujours une grande variabilité dans le nombre de clusters que les humains produisent pour le même ensemble de données. C'est pourquoi cette base de données a été renforcée par la confrontation entre différents partitionnements manuels faits par plusieurs participants, de manière à abaisser la subjectivité et avoir un bon ensemble de données pour l'évaluation.

3.4 CONCLUSION DU CHAPITRE

Dans ce chapitre nous avons introduit le concept de folksionary qui consiste en un dictionnaire regroupant les définitions d'un hashtag en unités de sens. Nous avons également défini un processus en quatres étapes pour construire le folksionary. D'abord, nous extrayons les définitions à partir des dictionnaires de hashtags disponibles en ligne, nous appliquons ensuite une distance sémantique pour mesurer l'écart entre les définitions pour chaque hashtag. Nous réalisons le partitionnement de définitions dans des unités de sens distinctes (clusters). Les clusters sont finalement présentées sous la forme d'un document pdf lisible par l'humain.

Nous avons également procédé à la validation de processus : nous avons développé une application web pour construire la vérité du terrain où les participants regroupent les définitions manuellement. Une évaluation pour comparer les résultats de notre processus de partitionnement automatique avec la vérité de terrain a été établie ensuite. Les résultats de l'évaluation montrent que notre approche fonctionne non seulement en théorie mais aussi dans la pratique : il fonctionne bien et donne de bons résultats en approchant la vérité du terrain de 89,2%. Nous avons aussi défini un processus permettant de découvrir les relations de synonymie entre les hashtags via un clustering hiérarchique.

Dans le chapitre qui suit nous utilisons le folksionary dans les traitements pour enrichir le profil d'un apprenant et afin de procéder à la recommandation des ressources pédagogiques à base de ce profil enrichi.

CHAPITRE 4

VERS UNE APPROCHE SOCIALE SEMANTIQUE POUR UNE RECOMMANDATION PERSONNALISEE DES CONTENUS PEDAGOGIQUES

Sommaire

Dans ce chapitre, nous présentons notre architecture globale pour une recommandation multidimensionnelle dans un environnement d'e-learning. Nous expliquons également notre méthode d'enrichissement du profil d'un apprenant à partir des hashtags contenus dans ses écrits au sein des réseaux sociaux.

4.1 APPROCHE GENERALE POUR UNE RECOMMANDATION MULTIDIMENSIONNELLE DANS UN ENVIRONNEMENT D'E-LEARNING

Aujourd'hui, les plates-formes d'e-learning contiennent un grand nombre de ressources, et offrent des cours pour les apprenants en fonction du niveau, des préférences, des intérêts, etc. Ces informations sont généralement recueillies à partir des renseignements fournis par les apprenants lors de la première inscription sur la plateforme, ou bien à travers l'historique de leurs activités et interactions avec le système. Toutefois, un grand nombre d'informations concernant l'apprenant ne sont pas inclus dans son profil d'apprentissage sur la plate-forme. Ce manque d'informations est dû au fait que les apprenants n'accordent que peu de temps et d'attention à alimenter les champs de leurs profil sur la plateforme où bien par manque de connaissance de leurs intérêts. En conséquence, de nombreuses ressources disponibles sur la plate-forme et intéressantes pour un apprenant donné ne lui sont pas proposées en raison du manque d'information sur son profil. Le but de notre travail est de définir une approche multidimensionnelle de recommandation qui permet aux apprenants d'accéder à plus de ressources que celles qui lui sont déjà proposées par la plateforme.

La recommandation est devenue un champ fertile de recherche ces dernières années. Elle est abordée dans divers domaines et applications, notamment le e-learning. Différentes approches ont été conçues comme nous l'avons présenté dans l'état de l'art (le filtrage collaboratif, le filtrage à base du contenu, le filtrage personnalisé, ou le filtrage sémantique). Cependant la plupart des contributions optent pour des approches hybrides pour des résultats optimaux de recommandation. Dans cette section nous présentons notre approche de recommandation multidimensionnelle dans un système d'e-learning fondée sur trois types de filtrage : personnalisé à base du

profil, social, et aussi à base des statistiques d'interaction avec le système. Nous nous focalisons ensuite sur la recommandation personnalisée à base du profil enrichi des activités d'un utilisateur sur les réseaux sociaux

Une des préoccupations majeures des systèmes d'e-learning d'aujourd'hui est d'améliorer leurs capacités en termes de définition et compréhension des profils de leurs adhérents [145, 165], en vue de leur recommander les meilleures ressources en fonction de leurs profils. En d'autres termes, mieux nous sommes informés sur le profil de l'apprenant mieux nous pouvons le servir. Or, la tâche de recommandation dans un environnement d'e-learning devient compliquée lorsqu'un nouvel apprenant arrive sur la plateforme directement après son inscription. Ceci est appelé le problème du démarrage à froid [166, 167] un des problèmes majeurs d'un système de recommandation. Dans cette perspective, diverses contributions comme cité dans la section 2.5.2 visent à obtenir plus d'informations sur le profil afin d'enrichir la connaissance des systèmes sur les apprenants pour une recommandation personnalisée. Cependant la plupart de ces contributions se fient aux informations données par un apprenant lors de son inscription ou bien déduites de son interaction avec le système [141, 168, 169]. Notre proposition est d'accroître le profil d'apprentissage en se basant sur les traces laissées par les apprenants sur les structures sociales, essentiellement les hashtags enrichis par leurs définitions. L'objectif est d'enrichir le champ des centres d'intérêts de l'apprenant en exploitant les données échangées et partagées par ce dernier sur les réseaux sociaux.

La figure 4.1 illustre notre architecture qui permet de lier les traces des utilisateurs consolidées dans le folksionary avec leurs profils de e-learning, et éventuellement de faire des recommandations multidimensionnelles des ressources pédagogiques.

Ci-après, nous donnons une description sur le rôle, les entrées et les résultats de chacune des trois couches de notre architecture à savoir : la couche de traitement des traces des apprenants, la couche d'enrichissement du profil et la couche de recommandation multidimensionnelle.

4.1.1 Traitement des Traces des Apprenants

Le travail dans cette couche est dédié à la collecte et au traitement des contributions des apprenants sur les réseaux sociaux. il s'agit d'une étape préparatoire des hashtags afin de les rendre utilisables pour les autres couches. D'un côté nous recueillons les hashtags pour chaque apprenant séparément pour construire sa per-

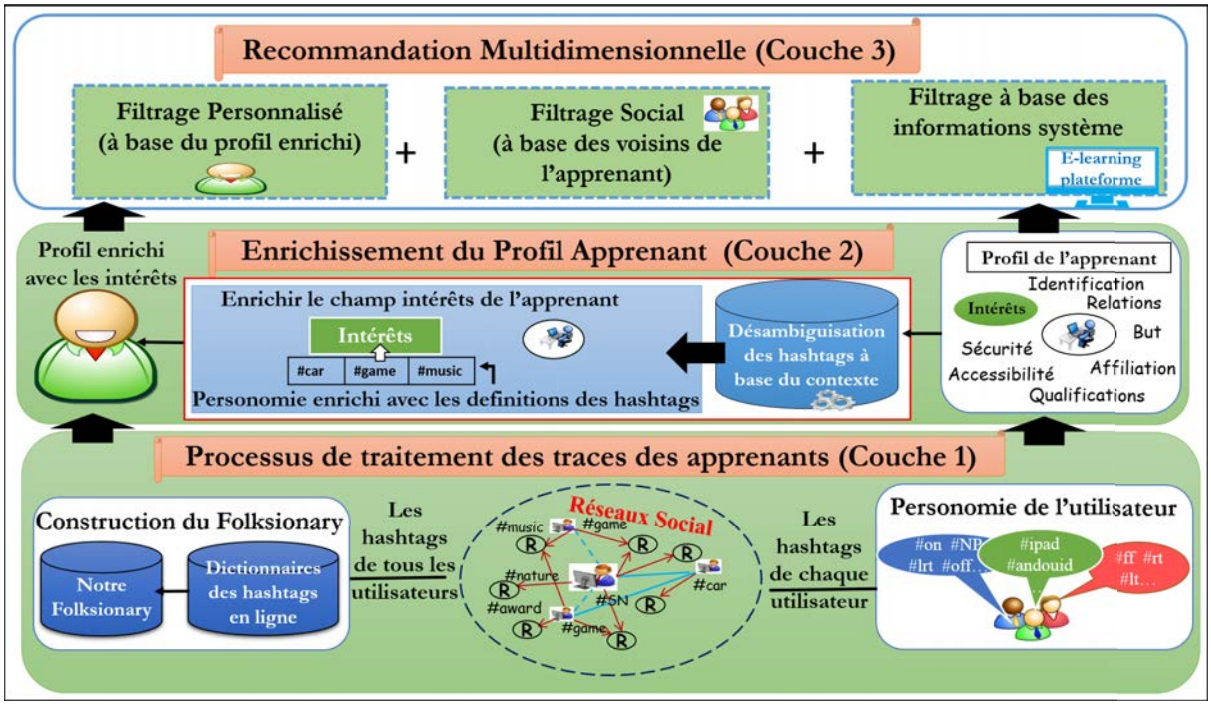

FIGURE *4.1: Architecture de recommandation multidimensionnelle dans les systèmes e-learning*

sonomie, d'autre part, nous explorons les hashtags utilisés sur les réseaux sociaux et leurs définitions sur les dictionnaires de définition de hashtags en ligne. Les définitions des hashtags sont ensuite consolidées dans un dictionnaire où ils sont désambiguïsés. Plus de détails sur cette étape sont donnés dans la section 3.2.

Dans cette phase nous avons deux traitement indépendants dont les résultats sont : (1) le folksionary qui consiste en un dictionnaire de hashtags (c.f. 3.2), qui à la manière d'un dictionnaire classique, associe à chaque terme (hashtag) un ensemble de sens, chaque sens pouvant avoir plusieurs définitions. (2) la personomie contenant les hashtags utilisées dans les données partagées par l'apprenant sur les réseaux sociaux ; ils fournissent une information sur les intérêts d'un apprenant.

4.1.2 Enrichissement du Profil Apprenant

Avant de procéder à l'enrichissement du profil de l'apprenant, nous avons modélisé son profil à base du standard LMS-LIP [129]. Nous effectuons ensuite l'enrichissement de ce profil avec les intérêts déduits de la personomie. Dans cette étape, nous faisons l'analyse nécessaire pour éliminer les ambiguïtés des hashtags selon leur contexte

d'utilisation dans le écrits des apprenants. Nous utilisons l'algorithme Lesk [153] pour lever l'ambiguïté des hashtags et leur affecter la meilleure définition en fonction du contexte d'utilisation. Ainsi les définitions produites représentent les intérêts d'un apprenant. Plus de détails sur cette étape seront données dans la section 4.2.

4.1.3 Recommandation Multidimensionnelle

L'objectif principal de ce travail est d'améliorer la recommandation dans les environnements d'e-learning, en proposant une approche qui intègre de nouvelles dimensions, à savoir : la dimension sociale, les annotations, les relations entre les profils, etc. Le modèle de recommandation est basé à la fois sur le profil apprenant enrichi et sur d'autres informations sociales pertinentes comme les voisins, ou les interactions de l'apprenant au sujet des cours sur les réseaux sociaux. D'autres informations importantes provenant directement de l'environnement d'e-learning peuvent également être incluses. Nous combinons trois types de filtrage (voir figure 4.2) : le filtrage personnalisé en fonction du profil, le filtrage social et le filtrage basé sur les statistiques du système.

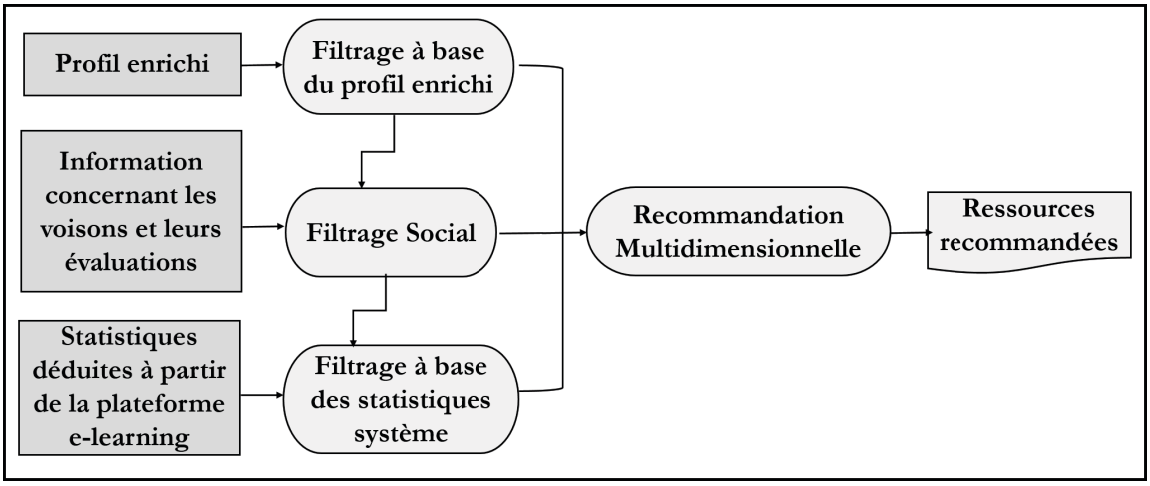

FIGURE *4.2: Modèle de recommandation multidimensionnelle*

Une vue synoptique du processus ci-après :
Données : Le profil enrichi, les ressources pédagogiques

1. **Filtrage à base du profil enrichi** (proposer à l'apprenant une sélection de ressources pédagogiques en fonction de son profil enrichi de manière automatique avec les centres d'intérêts)

- Calcul de la similarité entre les intérêts d'un apprenant (représentés par les définitions des hashtags dans le contexte de leur utilisation dans les écrits de l'apprenant) et la descriptions des ressources pédagogiques sur la plateforme d'e-learning.
- Sélection des ressources pédagogiques pertinentes, en cherchant dans leurs descriptifs celles comportant les centres d'intérêts de l'apprenant.

2. **Filtrage social** (se référer aux comportements passés des apprenants ayant des intérêts similaires pour recommander à l'apprenant sujet de la recommandation. Nous nous basons dans ce type de filtrage sur les relations existantes entre les hashtags : similarité, spécialisation ou généralisation etc.)
   - Pour chaque apprenant, calculer la similarité entre son profil et les profils de ses voisins.
   - Sélectionner les n profils les plus similaires au profil de l'utilisateur cible de la recommandation (Méthode du voisinage le plus proche [112]).
   - Faire des prédictions sur les ressources qui sont susceptibles d'intéresser l'utilisateur en utilisant à la fois les contenus similaires (Méthode à base de corrélation entre contenus dans le voisinage le plus proche [116]) à ceux déjà consultés et évalués par son voisinage le plus proche.
3. **Filtrage basé sur des statistiques système** (raffiner la liste de recommandations proposées en ayant recours aux évaluations des ressources sur la plateforme d'e-learning ex : si nous avons deux cours qui traitent le même sujet, proposer le mieux évalué ou bien le plus visité par les adhérents du système)
   - Sélectionner les ressources les plus visités par les apprenants.
   - Trier les ressources pertinentes pour un apprenant et sélectionner les mieux notées.
   - Recommander des ressources supplémentaires à l'apprenant en se basant sur l'histoire de ses interactions avec le système.

Résultats : l'ensemble des ressources à recommander

Dans ce qui précède nous avons présenté notre approche de recommandation multidimensionnelle dans un environnement d'e-learning. Nous nous focalisons sur la première dimension (filtrage à base du profil enrichi), les autres dimensions feront l'objet de travaux futurs. Dans la section 4.3 nous présentons notre algorithme de recommandation personnalisée à base du profil enrichi. Avant cela, nous présentons dans la section 4.2 le processus d'enrichissement des profils en utilisant les hashtags.

4.2 ENRICHISSEMENT DU PROFIL A PARTIR DES HASHTAGS

Dans cette section nous nous focalisons sur le processus de transformation des hashtags contenus dans les écrits d'un apprenant et leur conversion à un ensemble d'intérêts permettant d'enrichir son profil. La figure 4.3 illustre l'approche globale de ce processus étape par étape en commençant par la récupération des hahstags de l'apprenant sur des réseaux sociaux, ensuite le traitement pour les enrichir sémantiquement jusqu'à la génération d'un ensemble d'intérêts caractérisant l'apprenant et permettant d'avoir plus de connaissances sur ses besoins. Les sections suivantes donnent le détail de chacune de ces étapes.

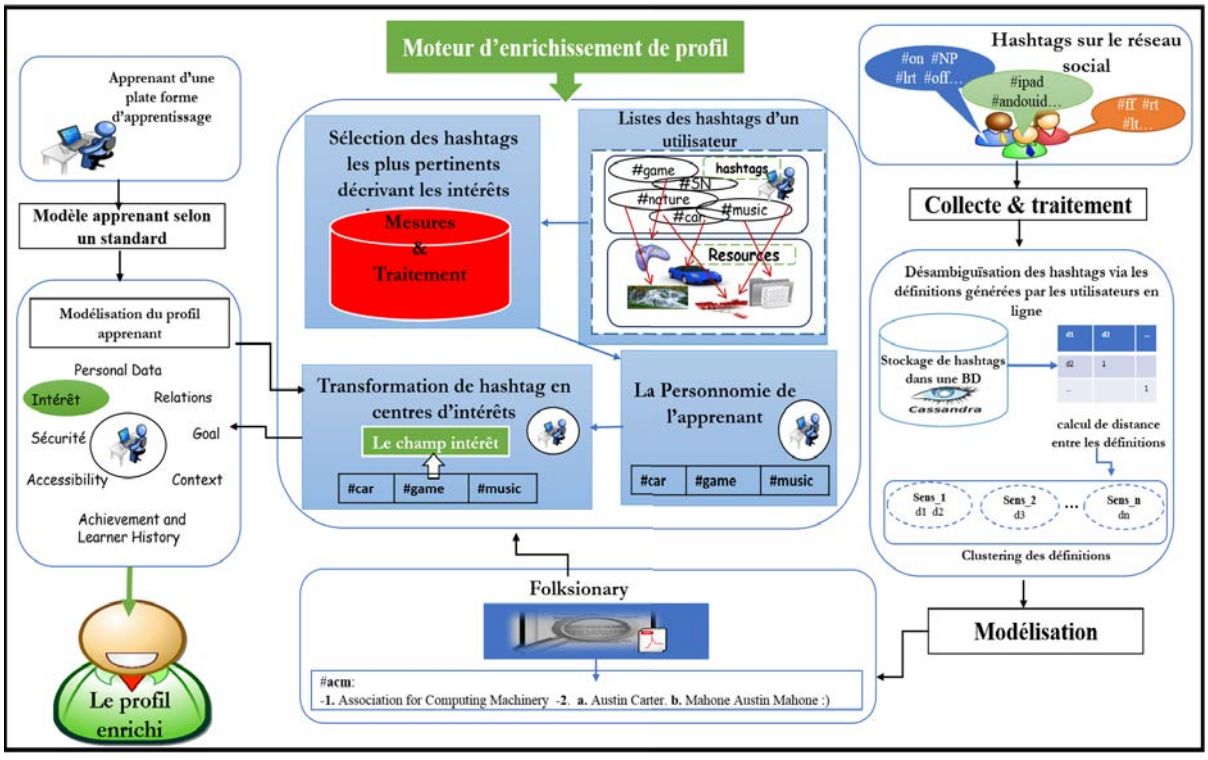

FIGURE *4.3: Approche d'enrichissement du profil apprenant*

4.2.1 Extraction de la personomie d'un apprenant

Les réseaux sociaux sont de riches structures de partage d'informations entre les utilisateurs, ils peuvent aussi être très influents quand il s'agit de construire une opinion, ou bien prendre des décisions etc. En effet, un apprenant participe activement sur les structures sociales et partage ainsi une quantité importante de données. Dans notre approche, nous nous concentrons sur le contenu des messages échangés sur les réseaux sociaux, nous sommes principalement intéressés à récupérer les hashtags contenus dans les données publiées par les apprenants. Ainsi, pour chaque apprenant nous collectons tous les hashtags qu'il utilise sur les réseaux sociaux, nous appelons cet ensemble de hashtags la personomie de l'apprenant. Dans ce qui suit, nous décrivons la méthode de construction de la personomie d'un apprenant (figure 4.4).

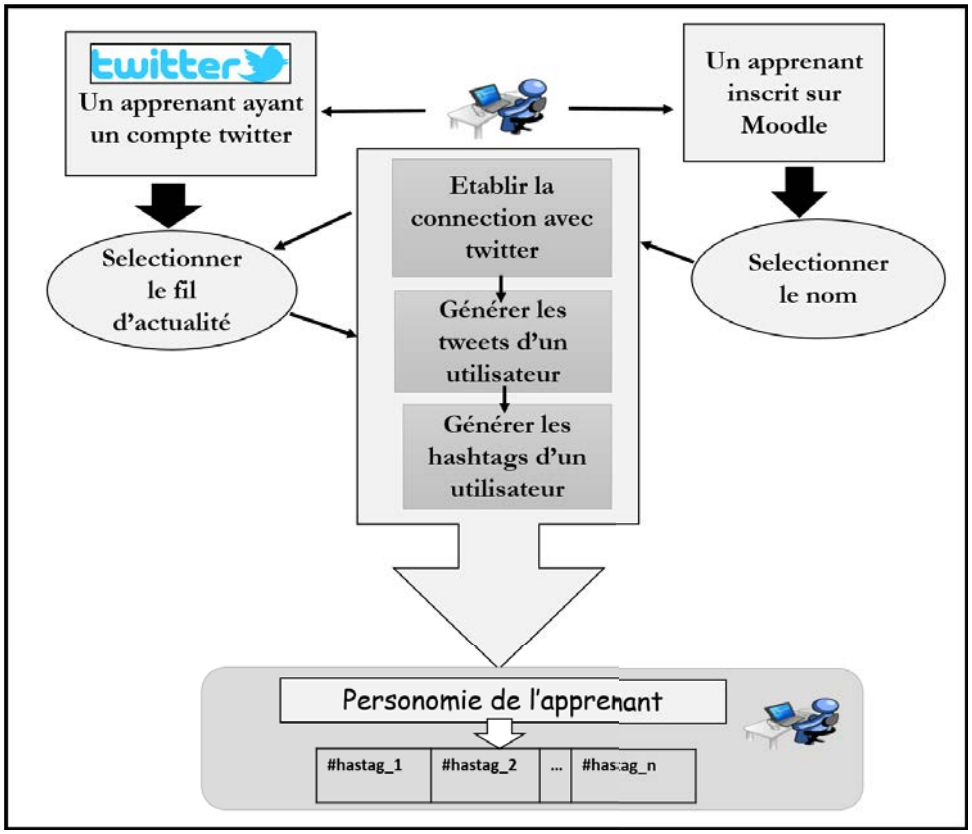

FIGURE *4.4: Construction de la personomie de l'apprenant à partir de Twitter*

Notre approche pour générer la personomie d'un apprenant est générique et peut être applicable à l'ensemble des réseaux sociaux où l'apprenant dispose d'un compte

et il est actif. Dans notre cas nous avons choisi de travailler avec Twitter [1] étant donné la richesse de son API qui fournit aux développeurs un ensemble de fonctionnalités très intéressantes. Il dispose d'une base de données complète, et des mesures quantitatives et qualitatives des discussions au sujet d'un hashtag, une personne, un compte, ou un ensemble de personnes. Il permet également la recherche dans l'historique qui est une caractéristique très importante pour notre travail.

Comme le montre bien la figure 4.4 la première étape est d'accéder au fil d'actualité d'un apprenant et récupérer les hashtags contenus dans les tweets. La dernière étape consiste à générer la personomie de l'apprenant. La figure 4.5 donne un exemple de personomie générée à partir de Twitter pour un utilisateur donné. Le programme permet d'abord de récupérer les tweets de l'apprenant et par la suite sélectionner l'ensemble de hashtags contenu dans ces derniers. Dans notre exemple la personomie est composée de quatre hashtags #betawards, #riseandshout, #avo et #acp. Or les hashtags collectés sont des entités ambigües. Afin de les utiliser pour des fins d'enrichissement, nous devons tout d'abord les désambiguïser. La section suivante est consacrée au processus de désambiguisation et de transformation de la personomie de l'apprenant en un ensemble d'intérêts.

4.2.2 Désambiguisation de la personomie et génération des intérêts

4.2.2.1 Désambiguïsation sémantique d'un mot

Les mots ambigus ont plusieurs sens, qui peuvent être liés ou non entre eux. Ces mots sont dits polysémiques. Ils peuvent être interprétés de différentes manières dans des contextes différents. La tâche qui consiste à définir le sens d'un mot ambigu en contexte se nomme Désambiguïsation lexicale pour Word Sense Disambiguation (WSD) en anglais. L'ensemble des sens candidats peuvent être dégagés à partir de bases de connaissances externes telles que les dictionnaires et les hiérarchies sémantiques (wordnet, wiktionary etc.) ou bien identifiés à l'aide des méthodes automatiques d'acquisition de sens. Tandis que le contexte est défini à travers les mots existants à gauche ou à droite du mot ambigu, les dépendances syntaxiques de ces derniers, les positions des mots, les traits pondérés etc. Une méthode de désambiguïsation permet d'attribuer un ou plusieurs sens aux instances des mots en contexte, il

1. http ://fr.wikipedia.org/wiki/Twitter

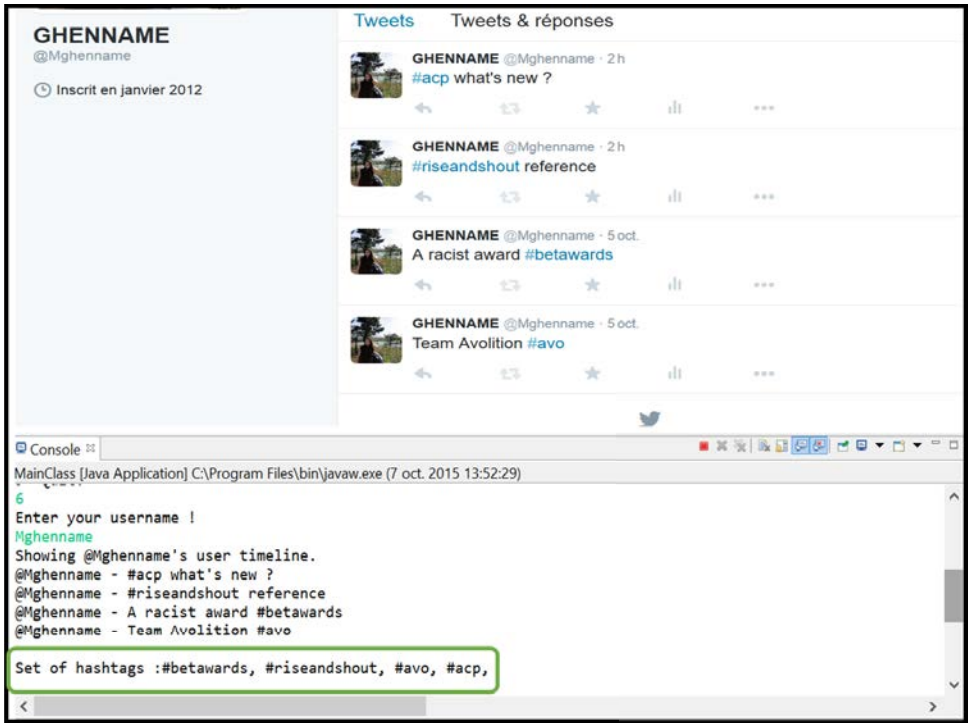

FIGURE *4.5: Exemple d'une personomie générée à partir Twitter*

existe plusieurs méthodes de désambiguïsation lexicale : les méthodes basées sur les connaissances et les méthodes basées sur les corpus-données.

A travers l'analyse des contextes dans lesquels les mots apparaissent dans les textes, il est possible d'identifier leur sens. Pour induire le sens il y a des méthodes qui se basent sur la construction de matrices de cooccurrence dont les cellules contiennent le score de similarité d'une paire de mots. D'autres travaux se basent sur la construction des vecteurs contextuels [170], ou encore les graphes de cooccurrence dont le principe est qu'un arc relie deux noeuds lorsque les mots apparaissent dans les mêmes contextes, sachant que les noeuds représentent les mots de ces noeuds qui cooccurrent avec le mot cible dans le corpus [171]. D'autres méthodes optent pour le clustering des vecteurs en fonction de leur similarité ou le partitionnement du graphe en régions de haute densité. Comme déjà mentionné précédemment, la seule manière d'identifier le sens d'un mot ambigu est de se référer à son contexte [172], ce qui fait que la nature et la quantité des informations contextuelles requises sont des notions importantes dans le travail de désambiguïsation. Ces deux notions dépendent du type d'ambigüité du mot, des relations existantes entre ses différents sens, ainsi que de la distinctivité et l'exclusion des sens lexicaux. En d'autres termes un mot pris séparé-

ment ne peut pas avoir un sens précis mais si on prend les N mots qui l'entourent et dès que N est assez grand il est possible de décider de manière non ambigu du sens du mot qu'on cherche à définir [173]. Différentes approches ont abordé le problème de désambiguïsation, ces approches peuvent être classées en quatre catégories :

Méthodes supervisées : Ces méthodes utilisent les informations sémantiques acquises directement à partir du texte. Elles sont connues pour être précises, et donnent les meilleurs résultats dans les évaluations des systèmes de désambiguïsation sémantique. Mais leur inconvénient majeur est la forte dépendance aux corpus de données pré-annotées dont l'élaboration est très coûteuse, et la difficulté d'avoir un ensemble d'entraînement annoté qui couvre tout le lexique d'une langue. Parmi les algorithmes les plus prédominants utilisés dans ce type de méthodes nous citons [174]. Ces algorithmes permettent de distinguer entre deux sens d'un mot ambigu à base de deux principales hypothèses : (1) la notion d'un sens par discours c'est à dire qu'un mot a le même sens tout au long d'un document, et (2) la notion d'un sens par collocation qui signifie que les mots qui se trouvent à proximité du mot ambigu constituent des indices fiables concernant son sens.

Méthodes non supervisées : Ces méthodes représentent une solution au problème de la disponibilité limitée de ressources. Elles sont intéressantes dans le sens où elles n'ont pas besoin de corpus sémantiquement annotés ou de sources externes de connaissances. Ici on distingue [175] : (1) des approches qui s'appuient sur la sortie des méthodes d'induction de sens à savoir le regroupement des instances des mots ambigus sur la base de leur similarité distributionnelle. (2) des approches à base de savoir qui utilisent des ressources lexicales et exploitent des connaissances linguistiques. (3) il y aussi des approches à base du clustering qui exploitent seulement les motifs présents dans les données, ainsi les clusters générés correspondent aux sens candidats des mots ambigus. Néanmoins, les méthodes non supervisées effectuent uniquement des comparaisons entre le sens des mots, et ont des difficultés à déterminer pourquoi et comment les significations des mots sont différentes. Parmi les algorithmes les plus utilisés dans ces méthodes, nous mentionnons [176].

Méthodes à base d'intelligence artificielle : Ces méthodes utilisent la fenêtre du mot cible (sac-de-mots entourant le mot cible) comme un élément important de la

désambiguïsation. Cependant la détérmination de la taille optimale de cette fenêtre reste compliquée et fait l'objet de plusieurs expériences [177]. Ces méthodes utilisent également un ensemble de corpus de données annoté crées manuellement pour former un algorithme, et ils sont très spécifiques au domaine. Parmi les algorithmes les plus importants de cette approche, nous mentionnons [178] et [179].

Méthodes basées sur les connaissances (exploitation des ressources prédéfinies, dictionnaires, thésaurus) : Dans ce type de méthodes, les ressources fournissent les descriptions sémantiques nécessaires pour faire la désambiguïsation, ainsi les différents sens décrits dans ces ressources sont les sens candidats parmi lesquels la méthode de désambiguïsation doit choisir pour définir le mot ambigu. Parmi les algorithmes les plus utilisés nous citons l'algorithme Lesk simplifié [180] qui repère toutes les définitions de sens du mot à désambiguïser, et ensuite définit le recouvrement entre définitions et nouveau contexte. Ainsi, le sens ayant le recouvrement le plus important avec le nouveau contexte est sélectionner comme sens du mot ambigu. Dans l'algorithme de Yarowsky [176], la désambiguïsation se fait en calculant le nombre de mots du contexte auxquels le thésaurus attribue un code thématique précis et le sens sélectionné est celui avec le score le plus élevé.

Nous avons opté pour une méthode basée sur les connaissances, étant donné que nous basons notre approche sur notre dictionnaire le foksionary et également Wordnet pour la désambiguïsation des sens des hashtags selon le contexte.

4.2.2.2 Désambiguïsation des hashtags d'un apprenant à partir du contexte

Notre objectif est de parvenir à définir le vrai sens dans lequel un utilisateur emploie ces hashtags à partir du contexte d'utilisation. Le hashtag peut être défini ou compris d'après la phrase ou l'ensemble du texte dans lesquels il se trouve, et la définition du hashtag sera l'objet de la discussion (tweet, statut, commentaire). Etant donné que le but final est d'enrichir le profil d'un apprenant et plus particulièrement le champ centre d'intérêt, on doit générer l'intérêt de l'utilisateur à partir du sens du hashtag. Un intérêt sera donc la définition attribué au hashtag à partir du contexte dans lequel il est abordé dans la discussion.

Comme déjà expliqué dans ce qui précède, la désambiguïsation d'un mot ambigu nécessite **un ensemble de sens candidats** pour le mot ambigu et **une méthode**

de désambiguïsation permettant de sélectionner le meilleur sens parmi les candidats. Dans notre cas (voir figure 4.6), le mot ambigu c'est le hashtag, tandis que les sens candidats sont l'ensemble des sens du hashtags dans le Folksionary. En ce qui concerne la méthode de désambiguïsation nous optons pour l'algorithme de Lesk [181] dont le calcul de similarité est extrêmement simple à calculer et ne requiert qu'un dictionnaire, tout en offrant une désambiguïsation de qualité raisonnable (50-70% de précision) [175]. De plus nous avons déjà validé son efficacité sur nos données lors de la construction du folksionary (c.f. section 3.3.2.4).

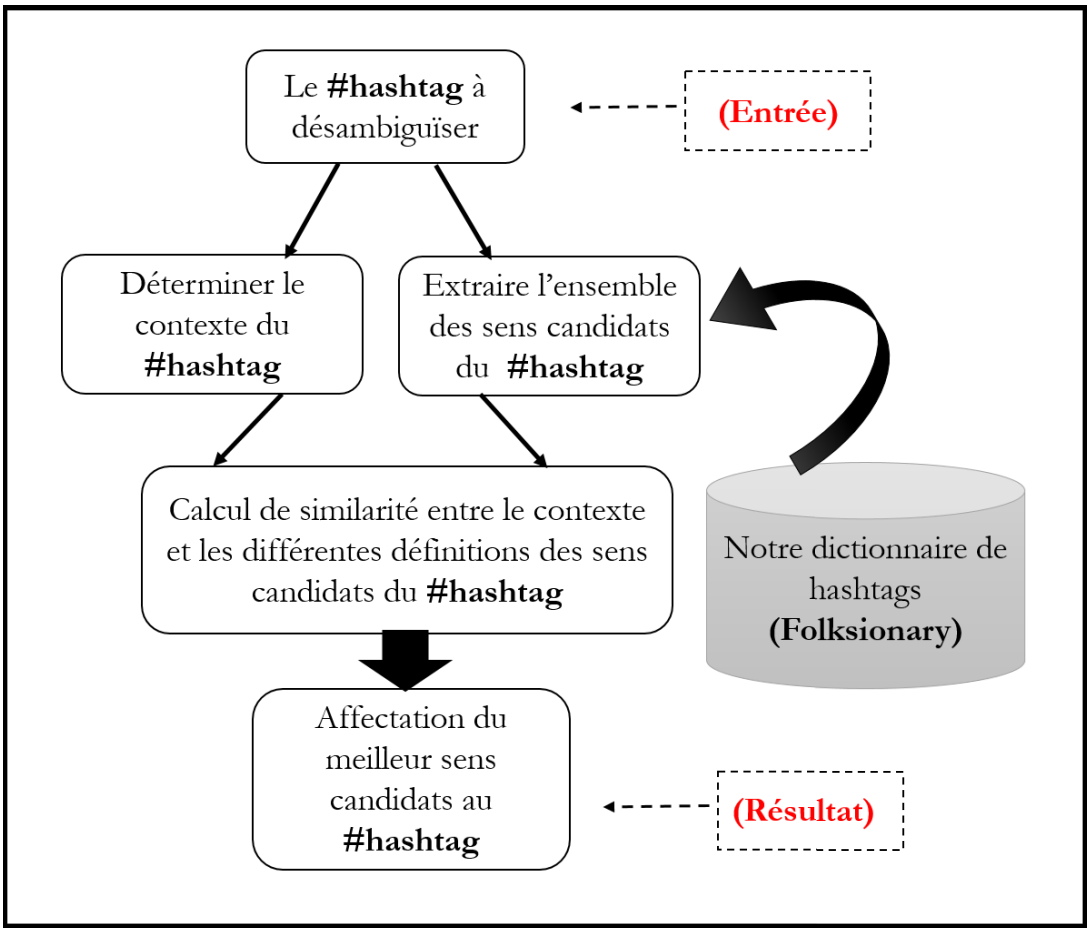

FIGURE *4.6: Processus de désambiguïsation de la personomie*

L'algorithme 4 présente les étapes de désambiguïsation de la personomie de l'apprenant à partir du contexte.

Ci-après nous donnons un exemple de sortie de sortie du processus de désambiguïsation de la personomie (figure 4.7).

$Personomie$: L'ensemble des hashtags à désambiguiser ;
$SensCandidats$: L'ensemble des sens candidats pour un hashtags ;
$Folksionary$: Dictionnaire des hashtags ;

pour *chaque hashtag t dans Personomie* **faire**
  $MeilleurScore = 0$;
  $MeilleurSensCandidat = sensCandidat(1)$;
  $Sup = 0$;
  Déterminer C(t) le contexte de t ;
  pour *chaque sensCandidat* s_j *dans SensCandidats* **faire**
    Extraire la définition $D(s_j)$ du Folksionary ;
    Calculer le nombre de superpositions sup = $D(s_j) \cap C(t)$;
    si $MeilleurScore \leq Sup$ **alors**
    fin
    $MeilleurScore = Sup$;
    $MeilleurSensCandidat = s_j$;
  fin
  Retourner(MeilleurSensCandidat) ;
fin

Algorithm 4: Désambiguïsation de la personomie de l'apprenant à partir de du contexte

4.2.3 Enrichissement des profils des apprenants

L'amélioration de l'expérience d'apprentissage dans les environnements d'e-learning est étroitement liée à la personnalisation des recommandations. En revanche pour offrir des ressources pertinentes aux apprenants, il est nécessaire d'avoir une bonne connaissance de leurs profils. Dans notre approche nous approchons le problème de recommandation personnalisée par un problème d'enrichissement du profil de l'apprenant en fonction de son activité au sein des réseaux sociaux. Dans ce travail, nous proposons d'enrichir le profil de l'apprenant à bases des hashtags désambiguïsés et enrichis par les définitions dans la section précédente. Cependant, avant de procéder à l'enrichissement, nous commençons par la modélisation du profil de l'apprenant. Il existe plusieurs méthodes de modélisation, mais comme notre travail est destiné à être une approche générique, nous avons opté pour une méthode normalisée pour la modélisation de l'apprenant.

Nous avons opté pour IMS Learner Information Package (IMS-LIP) [129], étant donnée sa capacité à faciliter l'échange d'informations sur les apprenants entre les

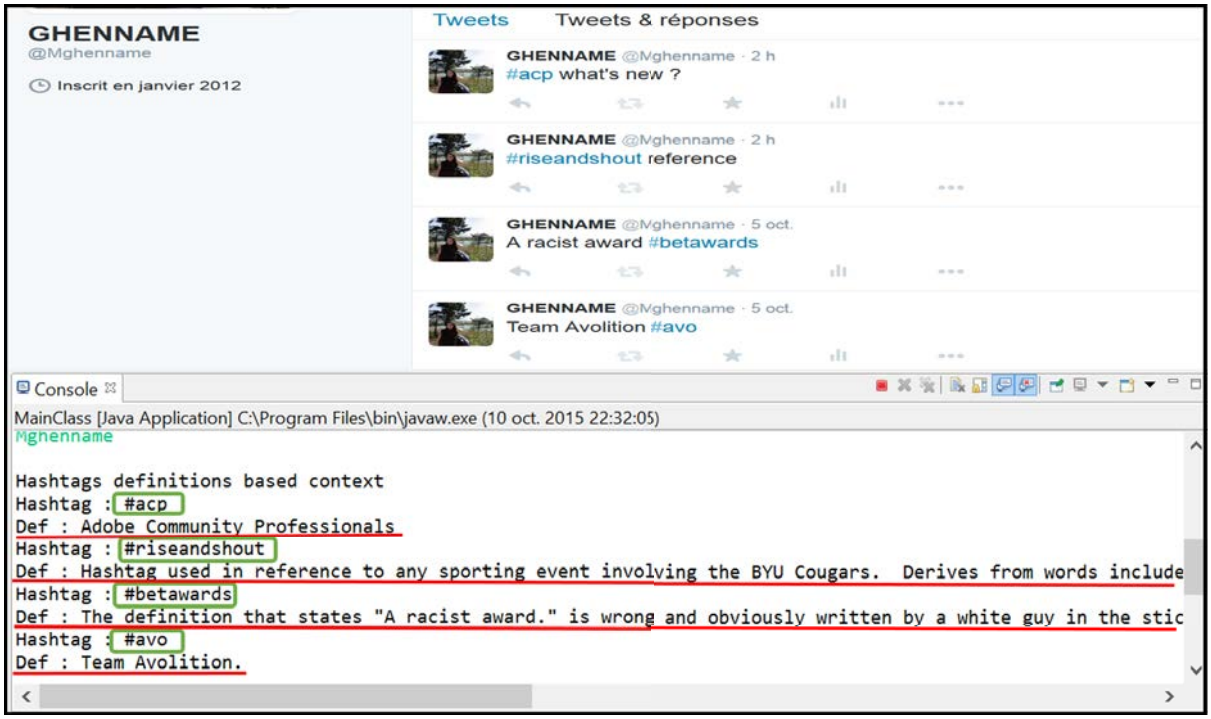

FIGURE *4.7: Exemple de désambiguïsation de la personomie et génération des intérêts*

systèmes d'éducation et les systèmes de gestion de l'apprentissage. IMS-LIP est un modèle de données défini en XML, il permet de décrire les caractéristiques des utilisateurs issues des utilisations générales telles que l'enregistrement et la gestion de l'historique, l'engagement de l'apprenant dans une expérience d'apprentissage, et permet de découvrir les possibilités d'apprentissage des apprenants [133]. IMS-LIP est structuré selon onze catégories de base [129] :

- **L'identification :** décrit les données démographiques et biographiques sur l'apprenant (exemple : nom, âge, adresse, email, etc.) ;
- **Le but :** définit l'objectif de la tâche d'apprentissage, l'attente de carrière et d'autres objectifs ;
- **Qualifications, Certifications & Licences (QCL) :** décrit l'ensemble des diplômes de l'apprenant ;
- **L'activité :** décrit toute activité liée à l'apprentissage dans n'importe quel état d'exécution (exemples : formation, expérience professionnelle, etc.) ;
- **Les intérêts :** maintiennent toutes les informations décrivant les hobbies de l'apprenant et les activités récréatives ;

- **les compétences :** décrit les compétences, l'expérience et les connaissances acquises ;
- **La transcription :** est utilisé pour fournir un résumé sur des résultats scolaires ;
- **L'affiliation :** présente des informations sur l'appartenance aux organisations professionnelles ;
- **L'accessibilité :** décrit l'accessibilité générale comme : les capacités linguistiques, les handicaps, les conditions d'admissibilité et les préférences d'apprentissage ;
- **La sécurité :** L'ensemble des mots de passe et clés de sécurité affectés à l'apprenant ;
- **Les relations :** décrit les relations entre les structures de données utilisées pour stocker les informations sur l'apprenant existant dans le modèle IMS-LIP.

Une fois la modélisation effectuée, nous procédons à l'enrichissent du profil de l'apprenant. Nous avons deux catégories d'informations (figure 4.6) (1) les informations issues directement de la plateforme d'e-learning à partir des champs renseignés par l'apprenant lors de sa première inscription, (2) le champs d'intérêts enrichi,avec les définitions de hashtags utilisés par l'apprenant dans ses écrits (section 4.2).

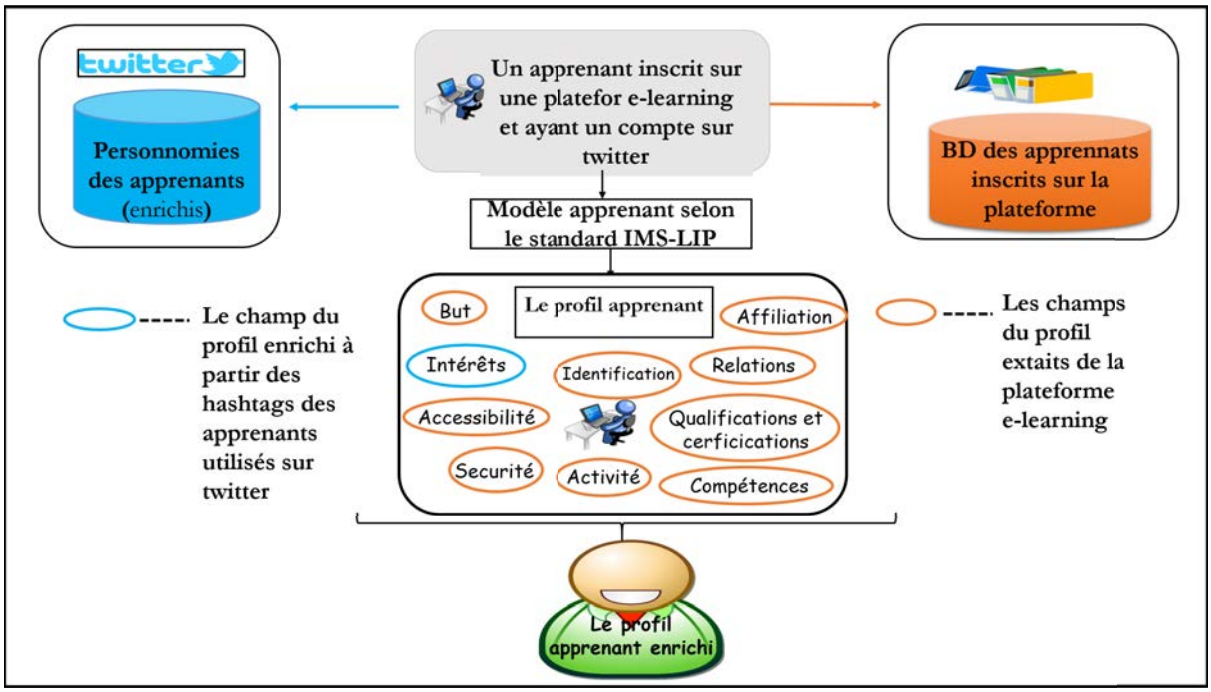

FIGURE *4.8: L'enrichissement du profil apprenant*

A la fin de cette étape, nous avons le profil de l'apprenant enrichi avec ses intérêts.

Ce profil enrichi peut être évolutif, car les intérêts d'un apprenant évoluent au fur et à mesure de l'évolution de ses activités sur les réseaux sociaux. En d'autres termes à chaque fois qu'un nouveau hashtag apparait dans les écrits d'un apprenant un traitement est effectué pour mettre à jour la personomie et par conséquent la liste des intérêts d'un apprenant. La personomie peut être également étendue grâce au clustering hiérarchique (c.f. section 3.2.2.4) qui permet de découvrir différentes relations entre les hashtags (synonymie, hyperonymie ou partie de etc.). Ainsi grâce à ces relations il est possible de découvrir d'autres intérêts de l'apprenant non explicitement exprimés dans sa personomie.

4.3 RECOMMANDATION DU CONTENU PEDAGOGIQUE DANS UN SYSTEME D'E-LEARNING

Ce travail de recherche porte sur la recommandation personnalisée dans les plateformes d'e-learning. Notre approche reste générique, elle considère comme données d'entrées un profil enrichi avec les hashtags et les descriptions de ressources pédagogiques sur une plateforme d'e-learning donnée. Etant donné que la modélisation du profil suit le standard IMS-LIP, la plupart des plates-formes d'e-learning savent construire des représentations des profils et contenus qui suivent ce format.

La figure 4.9 illustre le processus de calcul de distance entre les les intérêts de l'apprenant et les descriptions des cours existant sur la plate-forme, et la génération de l'ensemble des cours qui peuvent intéresser l'apprenant et qui ne lui sont pas proposés par le système.

Nous choisissons comme plateforme de test Moodle (Modular Objet-Oriented Dynamic Learning Environment) [2], une plateforme d'apprentissage destinée à fournir aux enseignants, administrateurs et apprenants un système unique robuste, sûr et intégré pour créer des environnements d'apprentissages personnalisés [182]. Ce choix s'appuie sur plusieurs facteurs que nous citons ci-après.

Pour commencer, Moodle est considéré lors de diverses évaluations comparatives avec des plateformes tel que Claroline, Dokeos, Ganesha, Sakai, WebCT et d'autres, comme la plus utilisée à travers le monde. Moodle a la confiance d'institutions et organisations grandes et petites, parmi lesquelles on compte Shell, la London School

2. https://moodle.org

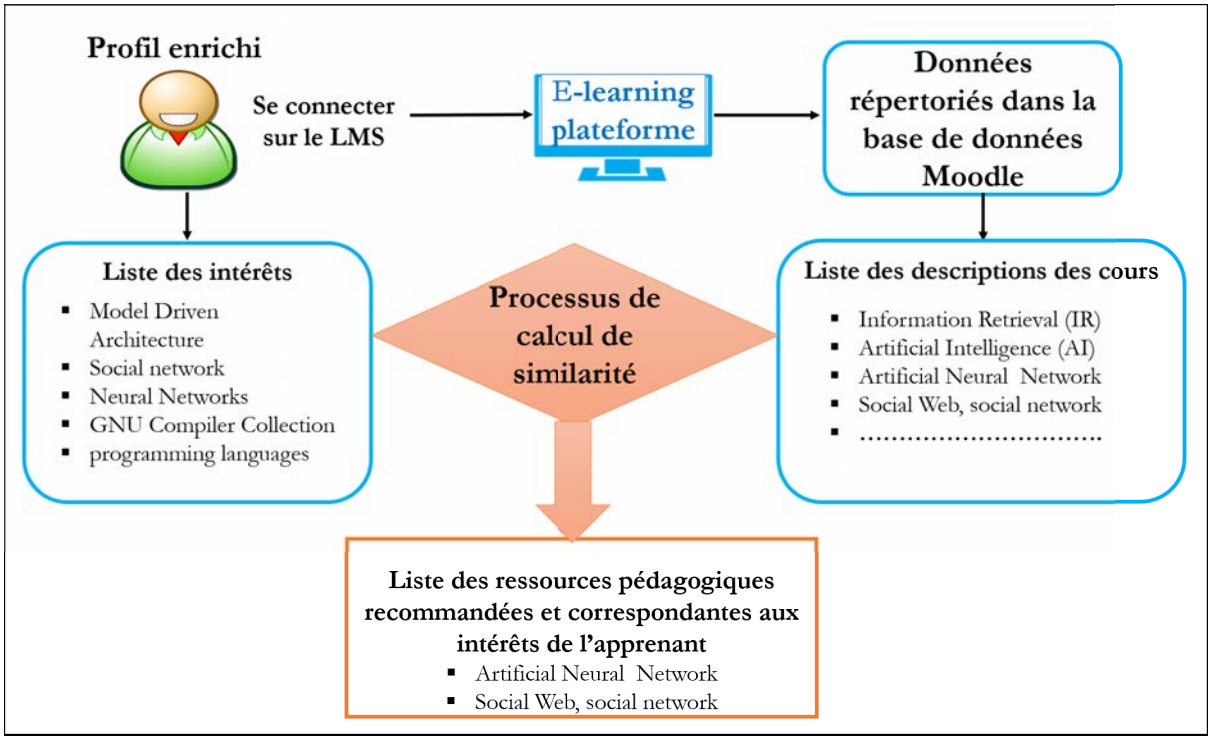

FIGURE *4.9: Recommandation personnalisée au sein d'une plateforme e-learning*

of Economics, l'Université d'État de New York (consulter *MoodleStats* et *MoodleE-tudeComparative*). Nous mentionnons aussi sa richesse en termes d'outils centrés sur l'apprenant et des environnements collaboratifs d'apprentissage qui renforcent tant l'enseignement que l'apprentissage. Sans oublier aussi qu'il est facile à utiliser, il maintient des registres détaillés de toutes les activités des utilisateurs, il est flexible et personnalisable, extensible, et contient plusieurs ressources disponibles en plusieurs langues. Toutes ces raisons font de Moodle une plateforme très utilisée que ce soit pour l'apprentissage ou même comme plateforme de test pour les chercheurs.

Un apprenant inscrit sur Moodle suit déjà un ensemble de cours, notre objectif est de proposer d'autres ressources qui ne lui sont pas proposés par le système et qui sont intéressants pour son profil. Moodle gère un ensemble d'informations décrivant chaque cours telle que : le nom du cours, la catégorie du cours, la date du début des cours, le numéro d'identification du cours, la description, la visibilité, etc. Après avoir collecté les descriptions des cours contenus dans la base de données Moodle, nous cherchons à faire correspondre la meilleur combinaison des cours aux intérêts d'un apprenant donné. Afin d'effectuer cette correspondance, nous avons défini un algorithme à base de clustering (voir algorithme 5) :

Données
SCR : l'ensemble des cours sur la plateforme e-learning
LER : l'apprenant avec un profil enrichi
interests : l'ensemble des intérêts d'un apprenant
threshold : le pas aléatoire généra par Markov pour le clustering

Result:
RecommendedRessources :
collection(String,collection(String,String))

$Sim_{LER,CR}$: float ;
ProposedCourses : collection(String,String) ;
SimilarityProcessing(String,String) ; calcule la distance entre la liste des descriptions des cours et les intérêts d'un apprenant
pour *chaque* $interest_i$ *dans interests* **faire**
  pour *chaque* CR_j *dans SCR* **faire**
    DC ← getDescription(CR_j) ;
    $Sim_{LER,CR}$ ← SimilarityProcessing($interest_i$,DC) ;
    si $Sim_{LER,CR} \leq$ *threshold* **alors**
      ProposedCourses.put(CR_j,DC) ;
      RecommendedRessources.put($interest_i$,ProposedCourses) ;
    fin
  fin
fin
retourner(RecommendedRessources) ;

Algorithm 5: Algorithme de recommandation à base du profil enrichi

4.4 Validation expérimentale avec Moodle

Dans cette section nous présentons les différents résultats générés par le programme de recommandation personnalisée à base du profil enrichi de l'apprenant dans une plate-forme d'e-learning.

La première étape, comme mentionné précédemment dans l'architecture générale de recommandation, est de faire le traitement des écrits laissés par l'apprenant sur ses réseaux sociaux dans le but de générer sa personomie. La figure 4.10 illustre

l'ensemble des intérêts de l'apprenant après enrichissement de son profil sur la plate-forme d'e-learning moodle, Les intérêts affichés sont le résultat de désambiguisation sémantique de l'ensemble de ses hashtags.

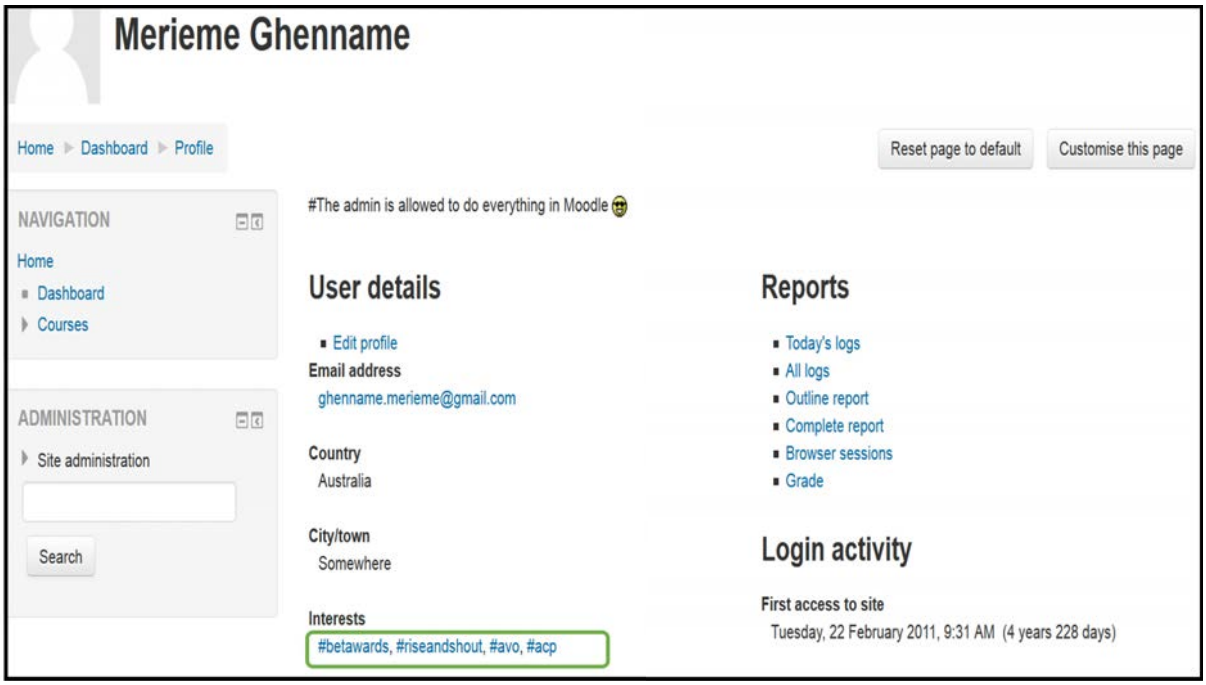

FIGURE *4.10: Test d'enrichissement de profil de l'apprenant avec le centre d'intérêt sur moodle*

L'intérêt est représenté sur le profil de l'apprenant dans la plate-forme d'e-learning par un label qui est le hashtag, et en cliquant sur ce dernier nous avons sa définition tirée du folksionnary. Cette définition est générée en fonction du contexte d'utilisation du hashtag dans les écrits de l'apprenant comme le montre bien la figure 4.11.

La figure 4.12 montre la liste des différents cous existant sur la plateforme d'e-learning, tous les cours possèdent des descriptions qui seront exploités dans le calcul de distance avec les intérêts générés précédemment afin de générer un ensemble de cours susceptibles d'intéresser un apprenant ayant des intérêts spécifiques.

La figure 4.13 illustre l'ensemble de cours calculé par notre algorithme de recommandation à base du profil enrichi 5. Les cours recommandés sont donc affichés sur le profil de l'apprenant lors de sa connexion à la plateforme moodle. La liste des recommandations est mise à jour à chaque fois que le profil de l'apprenant est mis à jour

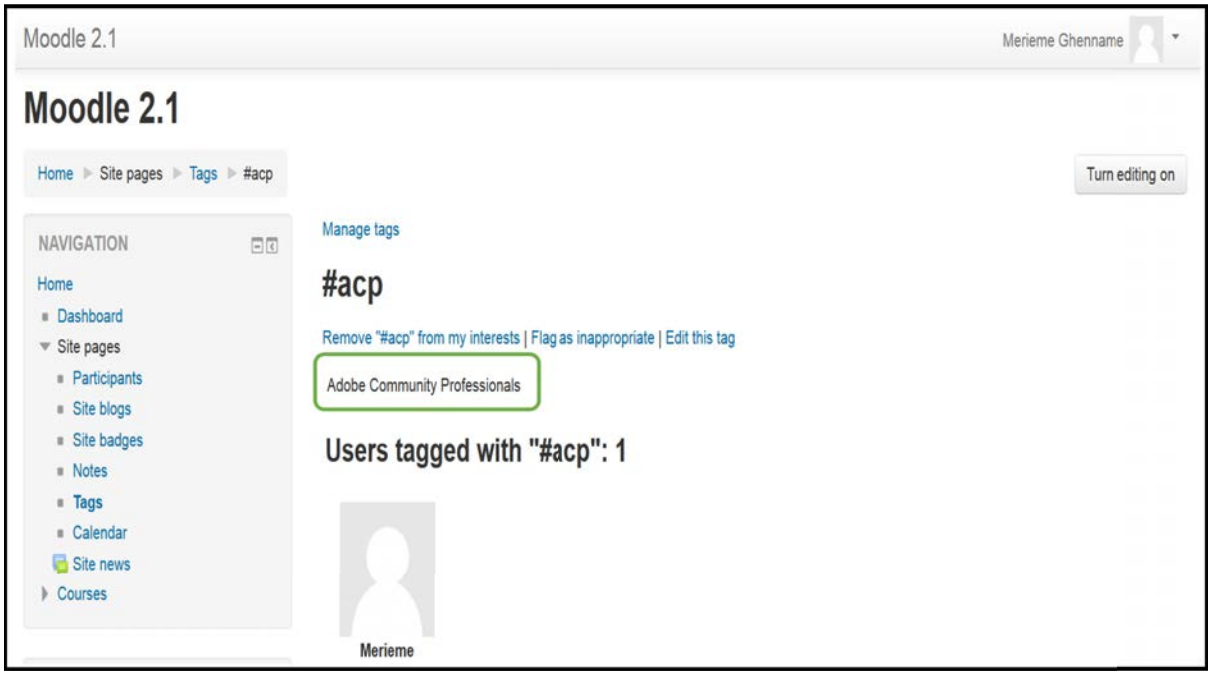

FIGURE *4.11: La définition explicite de l'intérêt de l'apprenant*

avec de nouveaux intérêts.

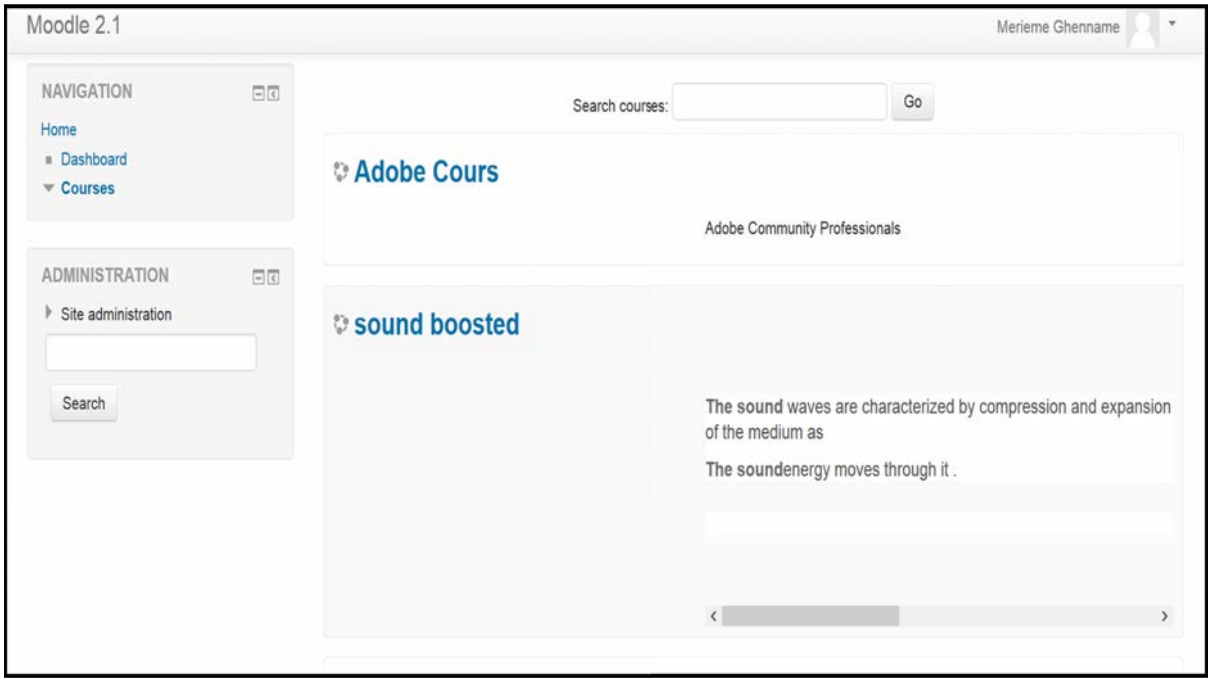

FIGURE *4.12: Un exemple de cours existants sur la plateforme moodle*

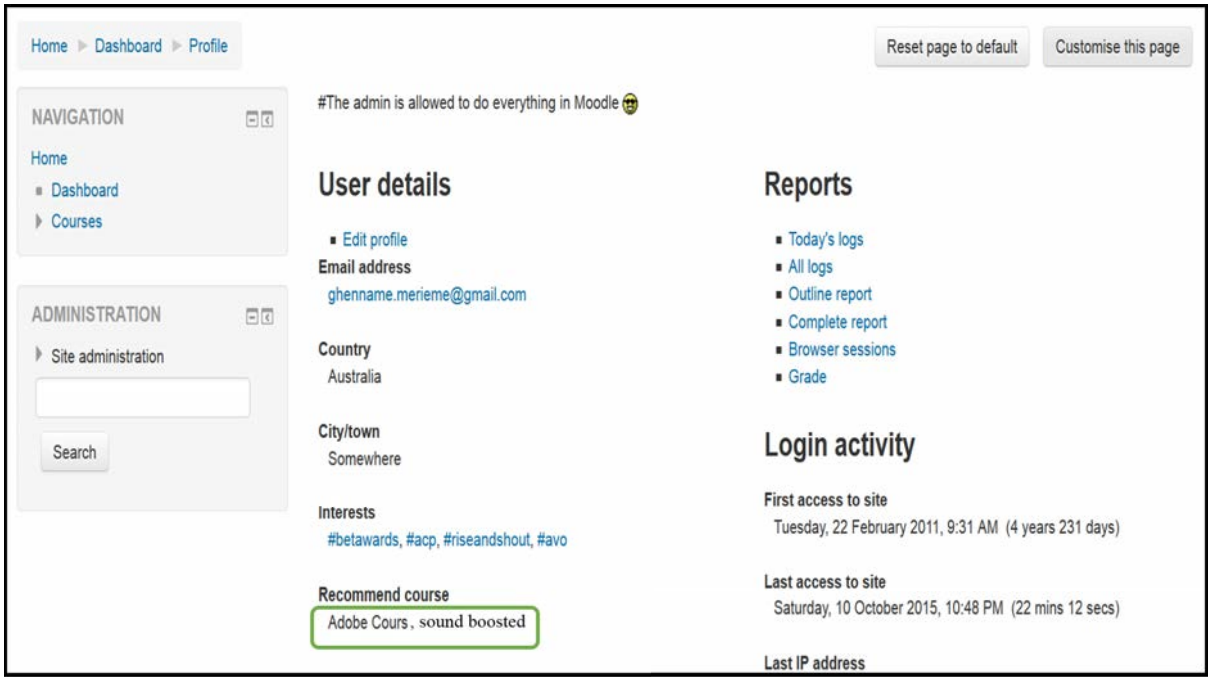

FIGURE *4.13: Liste des cours recommandés en fonction des intérêts de l'apprenant*

4.5 CONCLUSION DU CHAPITRE

Dans ce chapitre nous nous sommes consacrés au deuxième volet de notre travail qui consiste en la recommandation dans les systèmes e-learning. Nous avons proposé notre approche de recommandation multidimensionnelle à base de trois types de filtrage (Filtrage personnalisé à base du profil enrichi, filtrage social, et le filtrage à base des statistiques système). Dans ce chapitre nous avons exploité le travail effectué sur les hahstags dans le chapitre précédent dans le but d'enrichir le profil et ainsi nous nous sommes focalisés sur la recommandation personnalisée dans l'e-learning. Nous avons proposé un algorithme générique pour la recommandation personnalisée un environnement d'e-learning, et nous avons pris l'environnement Moodle comme plateforme pour tester notre approche (c.f.section 4.4).

CHAPITRE 5

CONCLUSION

Sommaire

5.1 CONTRIBUTION ET REGARD CRITIQUE

Dans l'introduction de cette thèse, nous avons présenté nos problématiques de recherche et les motivations de nos travaux tel que suit : Comment profiter de la dynamique participative du Web Social et les technologies de structuration et de formalisation du Web sémantique pour améliorer la recommandation dans un environnement d'e-learning ? Ainsi nous avons montré tout au long de ce travail quelles solutions et approches nous envisageons pour l'exploitation et la représentation des interactions et des activités issues du Web social à l'aide du Web sémantique, afin de produire des connaissances utiles pour des fins d'enrichissement des profils des apprenants et également de recommandation des ressources pédagogiques sur une plateforme d'e-learning. Nous revenons dans cette section sur les différentes réalisations et solutions apportées dans les deux volets majeurs (définis précédemment dans l'introduction) permettant la réalisation de nos travaux. Nous apportons ensuite un regard critique sur les résultats obtenus.

5.1.1 D'ANALYSE DES CONNAISSANCES SOCIALES POUR L'ENRICHISSEMENT DU PROFIL DE L'APPRENANT :

Nos travaux ici sont concentrés sur le traitement et l'analyse des traces des apprenants sur les réseaux sociaux, en particulier les hashtags. Notre but principal était de structurer et de donner un sens aux hashtags pour une future utilisation. Dans cette perspective nous avons introduit le concept de Folksionary qui consiste en un dictionnaire qui consolide les hashtags à la manière d'un dictionnaire classique qui associe à chaque hashtags un ensemble de sens, chacun pouvant contenir plusieurs définitions. Nous avons défini quatre étapes pour la construction du Folksionary : (1) dans un premier lieu nous collectons les hashtags les plus utilisées au sein des réseaux sociaux. (2) Ensuite nous appliquons une mesure de distance sémantique pour calculer la distance entre les définitions d'un hashtag. (3) Nous appliquons un clustering qui regroupe les définitions similaires dans des unités de sens distinctes pour chaque hashtag. (4) Et finalement nous formatons les clusters résultants sous forme d'un dictionnaire lisible par des humains. Nous avons également procédé à un clustering hiérarchique entres les sens des différents hashtags pour déduire les relations qui peuvent exister entre les hashtags. Et pour finir nous avons effectué une validation de nos résultats à travers la mesure de l'écart entre le Folksionary généré automatiquement et la vérité de terrain construite manuellement par un ensemble d'intervenants humains. Nous avons eu recours à la technique du Web scraping jScraper pour le recensement des définitions d'un hashtag et au classificateur de langue Apache Tika pour n'en garder que les définitions en anglais dans notre base de données. Nous avons opté pour Adapted Lesk [153] pour le calcul de distance sémantique entre les définitions, l'algorithme MCL [159] pour le clustering et l'algorithme CHA [183] pour le clustering hiérarchique. En ce qui concerne la mesure d'évaluation de la pertinente de notre folkionary nous avons choisi ACP 148. le choix des différents outils du Web sémantique et des techniques de data mining utilisées pour la construction du Folksionary ont été justifiés.

5.1.2 UNE APPROCHE SOCIALE SEMANTIQUE POUR UNE RECOMMANDATION PERSONNALISEE DES CONTENUS PEDAGOGIQUES :

Dans ce volet nous nous sommes intéressés d'une part à l'enrichissement du profil d'un apprenant en fonction des hashtags traités et analysés précédemment, et à la recommandation multidimensionnelle d'autre part. Nous avons choisi de suivre les activités des apprenants d'un environnement d'e-learning au sein des réseaux sociaux. Ainsi nous avons construit la personomie de l'apprenant contenant les hashtags utilisés dans les données qu'il partage sur les réseaux sociaux. Cette personomie est ensuite analysée dans le but de générer les intérêts d'un apprenant et d'enrichir son profil. La désambiguïsation de la personomie est effectuée avec l'algorithme Simplified Lesk [180]. Avant de passer à l'enrichissement du profil de l'apprenant nous modélisons ce dernier à base du standard IMS-LI [129]. Nous avons également présenté notre approche de recommandation multidimensionnelle à base de trois types de filtrage : (1) filtrage personnalisé à base du profil enrichi, (2) filtrage Social, et (3) filtrage à base de statistiques système. Nous nous sommes focalisés dans ce travail sur la première dimension et nous avons proposé un algorithme de clustering pour la recommandation personnalisée à base du profil enrichi (voir algorithme 5). Les approches et méthodes proposées dans ce travail restent générique et adaptable en fonction du réseau social et de la plateforme d'e-learning choisie. Dans ce travail nous avons choisi de suivre les activités d'un apprenant sur son réseau social Twitter et nous avons opté pour Moodle pour les raisons cités précédemment (c.f. section 4.3).

5.1.3 REGARD CRITIQUE :

Avant de conclure ce travail il est également important d'apporter un regard critique vis à vis des contributions dans le but d'ouvrir le chemin vers des travaux futurs. Ce travail explore le potentiel de cohabitation du Web social et des technologies du Web sémantique pour l'enrichissent du profil de l'apprenant à base de hashtags. Nous nous sommes concentrés sur l'enrichissent du champs centre d'intérêts de son profil, étant donné que les hashtags sont bien adaptés à cette information. Il aurait était sans doute très intéressant de combiner les hashtags avec d'autres types de contributions à savoir les conversations, les gens qu'il suit, les "j'aime" pour dégager d'autres informations utiles à l'enrichissent d'autres champs, comme les capacités, les buts etc. Il

va falloir aussi étendre notre méthode de récupération et traitement de hashtags dans le sens où à chaque fois qu'un nouveau hashtag apparait dans les dictionnaires en lignes, il doit être intégré directement dans le Folksionary. Enfin en termes d'évaluation, il serait pertinent de faire des tests durant un certain temps afin de relever des statistiques concernant l'engagement des apprenant et leur degré de satisfaction sur les recommandations avant et après l'instauration de notre approche.

5.2 TRAVAUX FUTURS

A l'issue de cette thèse, différentes perspectives de recherche dans la continuité de nos travaux s'offrent à nous. Nous résumons les travaux futurs principalement autour des problématiques suivantes :

- Découvrir d'autres relations que la synonymie entre les hashtags (hyperonymie, homonymie, antonymie, dérivés, etc.), qui nous permettront de mieux enrichir nos connaissances sur un apprenant et ainsi bien cerner ces besoins.
- Génération automatique d'une ontologie à partir du Folksionary, et ainsi créer une base lexicale à l'image de Wordnet permet de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical entre les hashtags. Nous appelons par convention cette base lexicale de hashtags TagNet.
- Mettre en oeuvre la totalité de l'approche de recommandation multidimensionnelle. Dans cette thèse nous avons proposé une approche de recommandation fondée sur trois types de filtrage : personnalisé à base du profil, social, et aussi à base des statistiques d'interactions avec le système, mais faute de temps nous nous somme focalisés sur la mise en oeuvre de la recommandation personnalisée à base du profil enrichi uniquement.
- Nous envisageons également de chercher à dégager d'autres informations à partir des hashtags combinées aux activités des apprenants afin de dégager d'autres connaissances capables d'enrichir d'autres champs du profil de l'apprenant à savoir par exemple les compétences, les buts, etc.

ANNEXE A

Activitées d'ouverture et d'innovation

- **Juin 2009 :** Membre du comité d'organisation de la conférence internationale des Réseaux et Service NGNs'9 à L'ENSIAS Rabat, Maroc.
- **Mars 2010 :** Responsable de la cellule de communication du comité d'organisation de la 1^{ere} édition du concours De programmation "Programming Moroccan Challenge" à l'ENSIAS Rabat, Maroc.
- **Mai 2010 :** Membre du comité d'organisation de la conférence internationale des systèmes intelligents SITA'10 à L'ENSIAS Rabat, Maroc.
- **Juillet 2010 :** Membre du comité d'organisation de la conférence internationale des Réseaux et Service NGNs'10 à Marrakech, Maroc.
- **Octobre 2011 :** Membre du comité d'organisation de JOIN-Med 2011 à L'ENSIAS Rabat, Maroc.
- **Avril 2012 :** Membre du comité d'organisation et chair de session durant la Conférence Global Engineering Education Conference (IEEE-EDUCON), à Marrakech, Maroc.
- **Novembre 2013 :** Membre du comité d'organisation du 3d. International Symposium ISKO-Maghreb'2013 (ISKO'13), Concepts and Tools for Knowledge Management(KM) à Marrakech, Maroc.
- **Juin 2014 :** Prix du 3^{eme} meilleur poster lors de la journée doctorale de la recherche autour du thème " Production scientifique du doctorant, évaluation, reconnaissance" à l'école nationale d'ingénieurs de Saint-Etienne.

Bibliographie

[1] A. Passant. *Technologies du web sémantique pour l'entreprise 2.0*. Ph.D. thesis, PhD thesis, Université Paris IV-Sorbonne (2009).

[2] T. Berners-Lee, J. Hendler, O. Lassila et al. *The semantic web*. Scientific american 284, 28 (2001).

[3] Y. Chabot. *Langage d'ontologies web owl 2 vue d'ensemble (deuxième édition)*. http ://yoan-chabot.fr/EspacePersonnel/Projet/W3C/owl2-overview/ (Mis en ligne en Juin 2014, consulté le 9 Octobre 2015).

[4] C. Farid. *Fouille de bases de données hétérogènes pour alimenter le web de données quel compromis entre qualité des données induites et robustesse des méthodes ?* In *IC 2009-23èmes journées francophones d'Ingénierie des Connaissances* (2012).

[5] M. HÉRIGAULT. *Moteur de recherche d'entreprise : déploiement du moteur sémantique exalead à la r&d de diagnostica stago*. Ph.D. thesis, MEMOIRE pour Titre professionnel, CONSERVATOIRE NATIONAL DES ARTS ET METIERS Ecole Management et Société (2012).

[6] I. Jilani and F. Amardeilh. *Enrichissement automatique d'une base de connaissances biologiques à l'aide des outils du web sémantique*. In *IC 2009-20èmes journées francophones d'Ingénierie des Connaissances* (2009), p. 47.

[7] M. Vernier and L. Monceaux. *Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques*. Traitement automatique des langues 51, 125 (2010).

[8] S. Fatiha. *Intégration sémantique de données guidée par une ontologie*. Ph.D. thesis, Thèse de Doctorat de l'Université Paris-Sud (2007).

[9] R. Parundekar, C. A. Knoblock and J. L. Ambite. *Linking and building ontologies of linked data*. In *The Semantic Web–ISWC 2010* (Springer, 2010), pp. 598–614.

[10] M. Baziz, N. Aussenac-Gilles and M. Boughanem. *Exploitation des liens sémantiques pour l'expansion de requêtes dans un système de recherche d'information.* In *INFORSID* (2003), pp. 121–134.

[11] H.-J. Kleebe and Y. D. Blum. *Sioc ceramic with high excess free carbon*. Journal of the European Ceramic Society 28, 1037 (2008).

[12] T. o'Reilly. *What is web 2.0* (O'Reilly Media, Inc., 2009).

[13] T. O'reilly. *What is web 2.0 : Design patterns and business models for the next generation of software*. Communications and Strategies 65, 17 (2007).

[14] L. Audet. *Wikis, blogues et web 2.0 : Opportunités et impacts pour la formation à distance*. Le Réseau d'enseignement francophone à distance du Canada (REFAD) (2010).

[15] F. Millerand, S. Proulx and J. Rueff. *Web social : mutation de la communication* (PUQ, 2010).

[16] J.-J. Richer. *Essai de définition du blog comme genre de discours*. Journées d'études : Les sites internet. Description et exploitation (2006).

[17] A. Degenne and M. Forsé. *Les réseaux sociaux*, volume 2 (Armand Colin Paris, 2004).

[18] G. Cormode and B. Krishnamurthy. *Key differences between web 1.0 and web 2.0*. First Monday 13 (2008).

[19] C. Sauvajol. *Big data et infobésité*. http ://www.bigdata-niort.fr/conference-infobesite-niort/ (Mis en ligne le 29 DECEMBRE 2010, consulté le 15 Mai 2014).

[20] M. Guy and E. Tonkin. *Tidying up tags*. D-lib Magazine 12, 1082 (2006).

[21] P. Mika. *Ontologies are us : A unified model of social networks and semantics*. In *The Semantic Web–ISWC 2005* (Springer, 2005), pp. 522–536.

[22] T. Vander Wal. *Folksonomy coinage and definition (2007)*. Retrieved fromhttp : llvanderwal. netlfolksonomy. html. Accessed : December (2009).

[23] B. Vickery. *La classification décimale universelle et l'indexage de la documentation technique*. Bulletin de l'Unesco à l'intention des bibliothèques 15, 57 (1961).

[24] M. Dewey and R. Couture-Lafleur. *Classification décimale de Dewey* (Groupe lyonnais de classification, 1970).

[25] T. Vander Wal. *Folksonomy*. online posting, Feb 7 (2007).

[26] E. Sutter. *Pour une écologie de l'information*. Documentaliste 35, 83 (1998).

[27] M. Baziz, M. Boughanem and N. Nassr. *La recherche d'information multilingue : désambiguisation et expansion de requêtes basées sur wordnet*. In *International Symposium On Programming and Systems (ISPS 2003)* (2003), pp. 175–186.

[28] A. MOUAKHER, M. DAOUD and S. YAHIA. *Folkoviz : Visualisation socio-sémantique de folksonomies*. ATELIER FOUILLE VISUELLE DE DONNEES : METHODOLOGIE ET EVALUATION p. 51 (2012).

[29] J. Caillouette, S. Garon, N. Dallaire, G. Boyer and A. Ellyson. *Étude de pratiques innovantes de développement des communautés dans les sept Centres de services de santé et de services sociaux de l'Estrie. Analyse transversale de sept études de cas* (Centre de recherche sur les innovations sociales, 2009).

[30] T. O'Reilly. *O'reilly spreading the knowledge of innovators*. What is web 2 (2005).

[31] J. Rogalski. *Le travail collaboratif dans la réalisation des tâches collectives*. L'intelligence pp. 147–159 (2005).

[32] C. Reffay and T. Chanier. *How social network analysis can help to measure cohesion in collaborative distance-learning*. In *Designing for change in networked learning environments* (Springer, 2003), pp. 343–352.

[33] A. Bandura and J.-A. Rondal. *L'apprentissage social*, volume 83 (P. Mardaga, 1980).

[34] J.-P. Pinte. *Vers des réseaux sociaux d'apprentissage en éducation*. Les Cahiers Dynamiques pp. 82–86 (2010).

[35] M. Linard. *Conception de dispositifs et changement de paradigme en formation*. Éducation permanente pp. 143–155 (2002).

[36] J.-P. Roux. *Socio-constructivisme et apprentissages scolaires* (2002).

[37] D. H. Jonassen. *Objectivism versus constructivism : Do we need a new philosophical paradigm ?* Educational technology research and development 39, 5 (1991).

[38] F. W. Mandroux. *L'apprentissage des langues à distance de l'ingénierie à la pédagogie*. centre de ressources et d'ingénierie documentaire (2011).

[39] V. de Poël, B. Lecomte et al. *Mooc, révolution ? business ? opportunité ?* (2013).

[40] A. KAVENOKY and G. TOUZOT. *La diffusion des technologies numériques dans la formation initiale et continue : Unit et utop*. 21ème Congrès Francais de Mécanique, 26 au 30 août 2013, Bordeaux, France (FR) (2013).

[41] B. Dorow and D. Widdows. *Discovering corpus-specific word senses*. In *Proceedings of the tenth conference on European chapter of the Association for*

Computational Linguistics-Volume 2 (Association for Computational Linguistics, 2003), pp. 79–82.

[42] A. L. Cabanac. *Confrontation à la perception humaine de mesures de similarité entre membres d'un réseau social académique*. In *Première conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique* (2010).

[43] P. Mercklé. *Sociologie des réseaux sociaux* (La découverte, 2011).

[44] S. Zhao, M. X. Zhou, Q. Yuan, X. Zhang, W. Zheng and R. Fu. *Who is talking about what : social map-based recommendation for content-centric social websites*. In *Proceedings of the fourth ACM conference on Recommender systems* (ACM, 2010), pp. 143–150.

[45] D. Poirier, F. Fessant, C. Bothorel, E. G. De Neef, M. Boullé et al. *Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films*. Revue des Nouvelles Technologies de l'Information (2009).

[46] M. Bank and J. Franke. *Social networks as data source for recommendation systems*. In *E-Commerce and Web Technologies* (Springer, 2010), pp. 49–60.

[47] F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso and F. Vernero. *Tag-based user modeling for social multi-device adaptive guides*. User Modeling and User-Adapted Interaction 18, 497 (2008).

[48] M. Szomszor, H. Alani, I. Cantador, K. O'Hara and N. Shadbolt. *Semantic modelling of user interests based on cross-folksonomy analysis*. In *The Semantic Web-ISWC 2008* (Springer, 2008), pp. 632–648.

[49] Z.-K. Zhang, C. Liu, Y.-C. Zhang and T. Zhou. *Solving the cold-start problem in recommender systems with social tags*. EPL (Europhysics Letters) 92, 28002 (2010).

[50] M. J. Barber. *Modularity and community detection in bipartite networks*. Physical Review E 76, 066102 (2007).

[51] T. Murata. *Detecting communities from tripartite networks*. In *Proceedings of the 19th international conference on World wide web* (ACM, 2010), pp. 1159–1160.

[52] K. Suzuki and K. Wakita. *Extracting multi-facet community structure from bipartite networks*. In *Computational Science and Engineering, 2009. CSE'09. International Conference on* (IEEE, 2009), volume 4, pp. 312–319.

[53] C. Bothorel. *Analyse de réseaux sociaux et recommandation de contenus non populaires*. Revue des nouvelles technologies de l'information (RNTI), A 5 (2011).

[54] I. Esslimani, A. Brun and A. Bayer. *Behavioral similarities for collaborative recommendations.* Journal of Digital Information Management 6 (2008).

[55] I. Esslimani, A. Brun and A. Boyer. *Enhancing collaborative filtering by frequent usage patterns*. In *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the* (IEEE, 2008), pp. 180–185.

[56] I. Esslimani, A. Brun, A. Boyer et al. *A collaborative filtering approach combining clustering and navigational based correlations.* In *WEBIST* (2009), pp. 364–369.

[57] I. Esslimani, A. Brun and A. Boyer. *From social networks to behavioral networks in recommender systems*. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in* (IEEE, 2009), pp. 143–148.

[58] I. Esslimani, A. Brun, A. Boyer et al. *Vers l'exploitation de la transitivité dans les réseaux comportementaux pour les systèmes de recommandations*. In *Intelligence collective et organisation des connaissances : 7ème colloque du chapitre français de l'ISKO-2009* (2009).

[59] I. Esslimani, A. Brun and A. Boyer. *Densifying a behavioral recommender system by social networks link prediction methods*. Social Network Analysis and Mining 1, 159 (2011).

[60] I. Esslimani, A. Brun and A. Boyer. *Detecting leaders in behavioral networks*. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on* (IEEE, 2010), pp. 281–285.

[61] I. Esslimani, A. Brun and A. Boyer. *Detecting leaders to alleviate latency in recommender systems*. In *E-Commerce and Web Technologies* (Springer, 2010), pp. 229–240.

[62] D. LE THRAN, P. Cheung-Mon-Chan and C. Bothorel. *Conception et développement de fonctionnalités innovantes liées à facebook pour un système de recommandation*. Rapport, Telecom Bretagne (2011).

[63] J. Golbeck, J. Hendler et al. *Filmtrust : Movie recommendations using trust in web-based social networks*. In *Proceedings of the IEEE Consumer communications and networking conference* (2006), volume 96, pp. 282–286.

[64] A. McNeill. *Trustedopinion. com offers netflix subscribers more relevant movie recommendations*. at, Sep 5 (2007).

[65] F. Carmagnola, F. Vernero and P. Grillo. *Sonars : A social networks-based algorithm for social recommender systems*. In *User Modeling, Adaptation, and Personalization* (Springer, 2009), pp. 223–234.

[66] C.-N. Ziegler and G. Lausen. *Analyzing correlation between trust and user similarity in online communities*. In *Trust management* (Springer, 2004), pp. 251–265.

[67] J. O'Donovan and B. Smyth. *Trust in recommender systems*. In *Proceedings of the 10th international conference on Intelligent user interfaces* (ACM, 2005), pp. 167–174.

[68] R. R. Sinha and K. Swearingen. *Comparing recommendations made by online systems and friends.* In *DELOS workshop : personalisation and recommender systems in digital libraries* (2001), volume 1.

[69] T. Olsson. *Decentralised social filtering based on trust*. In *proceedings of AAAI-98 Recommender Systems Workshop, Madison, WI* (1998).

[70] M. Montaner, B. López and J. L. de la Rosa. *Opinion-based filtering through trust*. In *Cooperative Information Agents VI* (Springer, 2002), pp. 164–178.

[71] J. Charlet, P. Laublet and C. Reynaud. *Web sémantique. rapport final de l'action spécifique 32, cnrs*. STIC (version 3 de décembre 2003), publié chez Cépadues (Hors-série de la collection Information interaction intelligence) (2003).

[72] T. R. Gruber. *Toward principles for the design of ontologies used for knowledge sharing ?* International journal of human-computer studies 43, 907 (1995).

[73] M. F. López, A. Gómez-Pérez, J. P. Sierra and A. P. Sierra. *Building a chemical ontology using methontology and the ontology design environment*. IEEE intelligent Systems 14, 37 (1999).

[74] A. Benayache. *Construction d'une mémoire organisationnelle de formation et évaluation dans un contexte e-learning : le projet memorae*. Ph.D. thesis, Compiègne (2005).

[75] W. W. W. Consortium et al. *Owl 2 web ontology language document overview* (2012).

[76] F. Baader and U. Sattler. *Description logics with aggregates and concrete domains*. Information Systems 28, 979 (2003).

[77] G. De Giacomo and M. Lenzerini. *Tbox and abox reasoning in expressive description logics.* KR 96, 316 (1996).

[78] S. J. Yang. *Context aware ubiquitous learning environments for peer-to-peer collaborative learning*. Educational Technology & Society 9, 188 (2006).

[79] Y. Prié and S. Garlatti. *Méta-données et annotations dans le web sémantique*. Revue I3 Information-Interaction-Intelligence 4, 45 (2004).

[80] P. Lévy. *L'intelligence collective : pour une anthropologie du cyberspace*, volume 11 (La Déouverte Paris, 1994).

[81] C. Cattuto, D. Benz, A. Hotho and G. Stumme. *Semantic grounding of tag relatedness in social bookmarking systems* (Springer, 2008).

[82] P. Mika. *Ontologies are us : A unified model of social networks and semantics*. In *The Semantic Web–ISWC 2005* (Springer, 2005), pp. 522–536.

[83] V. Tanasescu and O. Streibel. *Extreme tagging : Emergent semantics through the tagging of tags.* ESOE 292, 84 (2007).

[84] B. Huynh-Kim-Bang, E. Dané et al. *Social bookmarking et tags structurés.* Actes des 19es Journées Francophones d'Ingénierie des Connaissances (IC 2008) pp. 111–122 (2008).

[85] A. Monnin, F. Limpens, D. Laniado, F. L. Gandon et al. *L'ontologie nicetag : les tags en tant que graphes nommés*. Actes de l'atelier «Web Social», 10ième Journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010) (2010).

[86] A. Passant and P. Laublet. *Meaning of a tag : A collaborative approach to bridge the gap between tagging and linked data.* In *LDOW* (2008).

[87] L. Specia and E. Motta. *Integrating folksonomies with the semantic web*. In *The semantic web : research and applications* (Springer, 2007), pp. 624–639.

[88] T. Gruber. *Ontology of folksonomy : A mash-up of apples and oranges*. International Journal on Semantic Web and Information Systems (IJSWIS) 3, 1 (2007).

[89] M. Buffa, F. Gandon, G. Ereteo, P. Sander and C. Faron. *Sweetwiki : A semantic wiki*. Web Semantics : Science, Services and Agents on the World Wide Web 6, 84 (2008).

[90] H. Halpin, V. Robu and H. Shepherd. *The complex dynamics of collaborative tagging*. In *Proceedings of the 16th international conference on World Wide Web* (ACM, 2007), pp. 211–220.

[91] A. Sallaberry, N. Pecheur, S. Bringay, M. Roche and M. Teisseire. *Sequential patterns mining and gene sequence visualization to discover novelty from microarray data*. Journal of biomedical informatics 44, 760 (2011).

[92] F. Limpens, F. L. Gandon, M. Buffa et al. *Un cycle de vie complet pour l'enrichissement sémantique des folksonomies*. In *Proc. 11eme Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances EGC 2011, Hermann, Paris, Brest, France* (2011), pp. 1–12.

[93] R. T. Osguthorpe and C. R. Graham. *Blended learning environments : Definitions and directions.* Quarterly Review of Distance Education 4, 227 (2003).

[94] C. Raby, T. Karsenti, H. Meunier and S. Villeneuve. *Usage des tic en pédagogie universitaire : point de vue des étudiants*. Revue internationale des technologies en pépdagogie universitaire International Journal of Technologies in Higher Education 8, 6 (2011).

[95] M. Dagnaud. *Génération Y : les jeunes et les réseaux sociaux, de la dérision Ã la subversion* (CDE, 2013).

[96] S.-L. Huang and J.-H. Shiu. *A user-centric adaptive learning system for e-learning 2.0.* Journal of Educational Technology & Society 15 (2012).

[97] A. Bandura and J.-A. Rondal. *L'apprentissage social*, volume 83 (P. Mardaga, 1980).

[98] C. Costello. *Elgg 1.8 social networking* (Packt Publishing Ltd, 2012).

[99] S. Charmonman, R. Thirakomen and P. Mongkhonvanit. *Success stories of using social network in elearning*. Special Issue of the International Journal of the Computer, the Internet and Management 19 (2012).

[100] S. Tavales and S. Skevoulis. *Podcasts : Changing the face of e-learning.* In *Software Engineering Research and Practice* (2006), pp. 721–726.

[101] A. Anjomshoaa, K. V. Sao, A. M. Tjoa, E. Weippl and M. Hollauf. *Context oriented analysis of web 2.0 social network contents-mindmeister use-case*. In *Intelligent Information and Database Systems* (Springer, 2010), pp. 180–189.

[102] K. P. Brady, L. B. Holcomb and B. V. Smith. *The use of alternative social networking sites in higher educational settings : A case study of the e-learning benefits of ning in education.* Journal of Interactive Online Learning 9 (2010).

[103] J. Breslin, U. Bojars, A. Passant, S. Fernandez and S. Decker. *Sioc : Content exchange and semantic interoperability between social networks* (2009).

[104] R. Burke. *Hybrid recommender systems : Survey and experiments*. User modeling and user-adapted interaction 12, 331 (2002).

[105] M. J. Pazzani and D. Billsus. *Content-based recommendation systems*. In *The adaptive web* (Springer, 2007), pp. 325–341.

[106] C. Berrut and N. Denos. *Filtrage collaboratif*. Assistance intelligente à la recherche d'informations p. 30 (2003).

[107] R. Burke. *Hybrid recommender systems : Survey and experiments*. User modeling and user-adapted interaction 12, 331 (2002).

[108] T.-P. Liang, H.-J. Lai and Y.-C. Ku. *Personalized content recommendation and user satisfaction : Theoretical synthesis and empirical findings*. Journal of Management Information Systems 23, 45 (2006).

[109] A. Shepitsen, J. Gemmell, B. Mobasher and R. Burke. *Personalized recommendation in social tagging systems using hierarchical clustering*. In *Proceedings of the 2008 ACM conference on Recommender systems* (ACM, 2008), pp. 259–266.

[110] Mathieu. *Systéme de recommandation podcast science 83/les algorithmes de recommandation*. http ://www.podcastscience.fm/dossiers/2012/04/25/les-algorithmes-de-recommandation/ (Mis en ligne le 25 AVRIL 2012, consulté le 10 Octobre 2014).

[111] M. Balabanović and Y. Shoham. *Fab : content-based, collaborative recommendation*. Communications of the ACM 40, 66 (1997).

[112] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl. *Grouplens : an open architecture for collaborative filtering of netnews*. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (ACM, 1994), pp. 175–186.

[113] R. Hunt. *Percent agreement, pearson's correlation, and kappa as measures of inter-examiner reliability*. Journal of Dental Research 65, 128 (1986).

[114] D. T. Larose. *k-nearest neighbor algorithm*. Discovering Knowledge in Data : An Introduction to Data Mining pp. 90–106 (2005).

[115] D. Zhang and P. Bao. *Denoising by spatial correlation thresholding*. Circuits and Systems for Video Technology, IEEE Transactions on 13, 535 (2003).

[116] B. Sarwar, G. Karypis, J. Konstan and J. Riedl. *Item-based collaborative filtering recommendation algorithms*. In *Proceedings of the 10th international conference on World Wide Web* (ACM, 2001), pp. 285–295.

[117] P. Lops, M. De Gemmis and G. Semeraro. *Content-based recommender systems : State of the art and trends*. In *Recommender systems handbook* (Springer, 2011), pp. 73–105.

[118] J. J. Rocchio. *Relevance feedback in information retrieval* (1971).

[119] N. Béchet. *Etat de l'art sur les systemes de recommandation* .

[120] G. Adomavicius and A. Tuzhilin. *Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions*. Knowledge and Data Engineering, IEEE Transactions on 17, 734 (2005).

[121] P. Brusilovsky, A. Kobsa and W. Nejdl. *The adaptive web : methods and strategies of web personalization*, volume 4321 (Springer, 2007).

[122] D. Goldberg, D. Nichols, B. M. Oki and D. Terry. *Using collaborative filtering to weave an information tapestry*. Communications of the ACM 35, 61 (1992).

[123] D. Py. *Quelques méthodes d'intelligence artificielle pour la modélisation de l'élève*. Sciences et techniques éducatives 5 (1998).

[124] S. J. *Bypassing the intractable problem of student modeling*. Intelligent Tutoring Systems Conference, Montreal, Canada p. 1824 (1988).

[125] V. Dimitrova, G. Mccalla and S. Bull. *" open learner models : Future research directions" special issue of the ijaied (part 2)*. International Journal of Artificial Intelligence in Education 17, 217 (2007).

[126] P. Mendelsohn and P. Dillenbourg. *Le développement de l'enseignement intelligemment assisté par ordinateur*. In *Symposium Intelligence Naturelle et Intelligence Artificielle* (1991).

[127] W. Wahlster and A. Kobsa. *Dialogue-based user models*. Proceedings of the IEEE 74, 948 (1986).

[128] F. Farance. *Draft standard for learning technology. public and private information (papi) for learners (papi learner)*. Technical report, Version 6.0. Tech. Rep. Institute of Electrical and Electronics Engineers, Inc.. http ://ltsc. ieee. org/wg2/papi_learner_07_main. doc (2000).

[129] I. LIP. *Ims learner information package specification* (2008).

[130] I. RDCEO. *Ims reusable definition of competency or educational objective (rdceo)*. IMS GLC : Competencies Specification. Available at : http ://www. imsglobal. org/competencies (2002).

[131] I. G. L. Consortium et al. *Ims eportfolio specification*. Retrieved 1st of July (2006).

[132] L. Oubahssi and M. Grandbastien. *From learner information packages to student models : Which continuum ?* In *Intelligent Tutoring Systems* (Springer, 2006), pp. 288–297.

[133] L. Oubahssi. *Conception de plates-formes logicielles pour la formation à distance, présentant des propriétés d'adaptabilité à différentes catégories d'usagers et d'interopérabilité avec d'autres environnements logiciels*. Ph.D. thesis, Paris 5 (2005).

[134] R. Zghibi, S. Zghidi and O. Chater. *Les normes e-learning comme garant de qualité de l'enseignement à distance dans le contexte éducatif tunisien : le cas de l'uvt*. Frantice. net (2012).

[135] A. Behaz and M. Djoudi. *Approche de modélisation d'un apprenant à base d'ontologie pour un hypermédia adaptatif pédagogique.* In *CIIA* (2009).

[136] A. BAGHLI. *Diag-k : Ontologie pour l'élaboration du modèle de connaissances de l'apprenant dans les eiah* (2012).

[137] O. Sanjuan-Martinez, B. C. P. G-Bustelo, R. G. Crespo and E. T. Franco. *Using recommendation system for e-learning environments at degree level*. International Journal of Interactive Multimedia and Artificial Intelligence 1 (2009).

[138] V. Butoianu, O. Catteau, P. Vidal and J. Broisin. *Un systéme à base de traces pour la recherche personnalisée d'objets pédagogiques : le cas d'ariadne finder*. Atelier : (Personnalisation de l'apprentissage : quelles approches pour quels besoins ?), EIAH 2011 (2011).

[139] L. Berkani, A. Chikh and O. Nouali. *Recommandation personnalisée des ressources dans une communauté de pratique de e-learning. une approche à base de filtrage hybride*. In *INFORSID* (2013), pp. 131–138.

[140] R. Zouag. *Filtrage collaboratif des objets pédagogiques.* Ph.D. thesis (2014).

[141] C.-M. Chen, H.-M. Lee and Y.-H. Chen. *Personalized e-learning system using item response theory*. Computers & Education 44, 237 (2005).

[142] A. Klašnja-Milićević, B. Vesin, M. Ivanović and Z. Budimac. *E-learning personalization based on hybrid recommendation strategy and learning style identification*. Computers & Education 56, 885 (2011).

[143] A. BELHABIB and M. A. MATAHRI. *Conception d'un système de recommandation pour un réseau sociale d'apprentissage.* Ph.D. thesis (2014).

[144] M. Mullenweg, R. Boren, M. Jaquith, A. Ozz and P. Westwood. *Wordpress* (2011).

[145] M. B. ^{2}Florence Sèdes. *Modélisation basée sur des ontologies pour développer des recommandations personnalisées dans les systèmes hypermédia adaptatifs*. In *INFORSID* (2010), pp. 61–75.

[146] A. Purandare and T. Pedersen. *Discriminating among word meanings by identifying similar contexts*. In *AAAI* (2004), pp. 964–965.

[147] J. Ramos. *Using tf-idf to determine word relevance in document queries*. In *Proceedings of the first instructional conference on machine learning* (2003).

[148] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis and E. E. Milios. *Semantic similarity methods in wordnet and their application to information retrieval on the web*. In *Proceedings of the 7th annual ACM international workshop on Web information and data management* (ACM, 2005), pp. 10–16.

[149] V. N. Gudivada, V. V. Raghavan, W. I. Grosky and R. Kasanagottu. *Information retrieval on the world wide web*. IEEE Internet Computing pp. 58–68 (1997).

[150] C. D. Manning, P. Raghavan, H. Schütze et al. *Introduction to information retrieval*, volume 1 (Cambridge university press Cambridge, 2008).

[151] Y. Matsuo and M. Ishizuka. *Keyword extraction from a single document using word co-occurrence statistical information*. International Journal on Artificial Intelligence Tools 13, 157 (2004).

[152] S. Patwardhan, S. Banerjee and T. Pedersen. *Using measures of semantic relatedness for word sense disambiguation*. In *Computational linguistics and intelligent text processing* (Springer, 2003), pp. 241–257.

[153] S. Banerjee and T. Pedersen. *An adapted lesk algorithm for word sense disambiguation using wordnet*. In *Computational linguistics and intelligent text processing* (Springer, 2002), pp. 136–145.

[154] M. Lesk. *Information in data : using the oxford english dictionary on a computer*. In *ACM SIGIR Forum* (ACM, 1986), volume 20, pp. 18–21.

[155] A. Budanitsky and G. Hirst. *Evaluating wordnet-based measures of lexical semantic relatedness*. Computational Linguistics 32, 13 (2006).

[156] S. Khuller, S. G. Mitchell and V. V. Vazirani. *On-line algorithms for weighted bipartite matching and stable marriages*. Theoretical Computer Science 127, 255 (1994).

[157] H. W. Kuhn. *The hungarian method for the assignment problem*. Naval research logistics quarterly 2, 83 (1955).

[158] S. Brohee and J. van Helden. *Evaluation of clustering algorithms for protein-protein interaction networks*. BMC bioinformatics 7, 488 (2006).

[159] A. J. Enright, S. Van Dongen and C. A. Ouzounis. *An efficient algorithm for large-scale detection of protein families*. Nucleic acids research 30, 1575 (2002).

[160] F. Murtagh. *A survey of recent advances in hierarchical clustering algorithms*. The Computer Journal 26, 354 (1983).

[161] Y. Zhao and G. Karypis. *Evaluation of hierarchical clustering algorithms for document datasets*. In *Proceedings of the eleventh international conference on Information and knowledge management* (ACM, 2002), pp. 515–524.

[162] E. Hatcher, O. Gospodnetic and M. McCandless. *Lucene in action* (2004).

[163] T. Abeel, Y. Van de Peer and Y. Saeys. *Java-ml : A machine learning library*. The Journal of Machine Learning Research 10, 931 (2009).

[164] M. Burset and R. Guigo. *Evaluation of gene structure prediction programs*. Genomics 34, 353 (1996).

[165] K. I. B. Ghauth and N. A. Abdullah. *Building an e-learning recommender system using vector space model and good learners average rating*. In *Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on* (IEEE, 2009), pp. 194–196.

[166] R. Lara, D. Olmedilla, S. Arroyo, H. Lausen, D. Roman and P. Chirita. *A semantic web services framework for distributed e-learning environments*. Informe técnico, L3S Research Center (2004).

[167] M. W. Chughtai, A. Selamat, I. Ghani and J. J. Jung. *E-learning recommender systems based on goal-based hybrid filtering*. International Journal of Distributed Sensor Networks 2014 (2014).

[168] A. Zapata, V. H. Menéndez, M. E. Prieto and C. Romero. *Evaluation and selection of group recommendation strategies for collaborative searching of learning objects*. International Journal of Human-Computer Studies 76, 22 (2015).

[169] R. Zouag. *Filtrage collaboratif des objets pédagogiques.* Ph.D. thesis (2014).

[170] H. Schütze. *Automatic word sense discrimination*. Computational linguistics 24, 97 (1998).

[171] C. Benzitoun, E. Campione, J. Deulofeu, S. Henry, F. Sabio, S. Teston, A. Valli and J. Véronis. *L'analyse syntaxique de l'oral : problèmes et méthodes*. In *journée d'étude :" méthodes et outils pour l'évaluation des analyseurs syntaxiques"; organisée par l'Association pour le Traitement Automatique des LAngues (ATALA)* (2004), pp. 1–8.

[172] N. Ide and J. Véronis. *Introduction to the special issue on word sense disambiguation : the state of the art*. Computational linguistics 24, 2 (1998).

[173] C. E. Shannon and W. Weaver. *The mathematical theory of information* (1949).

[174] W. A. Gale, K. W. Church and D. Yarowsky. *A method for disambiguating word senses in a large corpus*. Computers and the Humanities 26, 415 (1992).

[175] A. Tchechmedjiev. *Etat de l'art mesures de similarité sémantique locales et algorithmes globaux pour la désambiguisation lexicale à base de connaissances*. In *Actes de la conférence conjointe JEP-TALNRECITAL* (2012), pp. 295–308.

[176] D. Yarowsky. *Unsupervised word sense disambiguation rivaling supervised methods*. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (Association for Computational Linguistics, 1995), pp. 189–196.

[177] J. Preiss and M. Stevenson. *Introduction to the special issue on word sense disambiguation*. Computer Speech & Language 18, 201 (2004).

[178] S. L. Small. *Word expert parsing : A theory of distributed word-based natural language understanding* (1980).

[179] K. Dahlgren. *Naive semantics for natural language understanding* (Springer, 1988).

[180] F. Vasilescu and P. Langlais. *Désambiguisation de corpus monolingues par des approches de type lesk*. Ph.D. thesis, Université de Montréal (2003).

[181] A. Kilgarriff and J. Rosenzweig. *English senseval : Report and results.* In *LREC* (2000).

[182] *A propos de moodle*. https ://docs.moodle.org/2x/fr/ Consulté le : 21-06-2015.

[183] M. Bruynooghe. *Classification ascendante hiérarchique des grands ensembles de données : un algorithme rapide fondé sur la construction des voisinages réductibles*. Les cahiers de l'analyse des Données 3, 7 (1978).

[184] P. L. SALAM and V. VALMAS. *Etude comparative des compétences développées dans deux formations hybrides de tuteurs en ligne : interactions à distance en asynchrone pour l'une et en synchrone pour l'autre*. In *Actes du colloque Epal 2009 (Echanger pour apprendre en ligne : conception, instrumentation, interactions, multimodalité), université Stendhal - Grenoble* (2009).

[185] A. Mounia. *From e-learning to e-learning 2.0 : What impact on the quality of learning*. The Journal of Quality in Education p. 12 (2012).

[186] R. Peirano. *"les pédago-blogueurs"*. http ://skolanet.over-blog.fr/article-les-pedago-blogueurs-synthese-41997719.html (Mis en ligne le 29 DECEMBRE 2010, consulté le 23 septembre 2014).

[187] L. Audet. *Wikis, blogues et web 2.0 : Opportunités et impacts pour la formation à distance*. Le Réseau d'enseignement francophone à distance du Canada (REFAD) (2010).

[188] T. Murata and S. Moriyasu. *Link prediction based on structural properties of online social networks*. New Generation Computing 26, 245 (2008).

[189] C. Bothorel. *Analyse de réseaux sociaux et recommandation de contenus non populaires*. Revue des nouvelles technologies de l'information (RNTI), A 5 (2011).

[190] C. Bizer, T. Heath, K. Idehen and T. Berners-Lee. *Linked data on the web (ldow2008)*. In *Proceedings of the 17th international conference on World Wide Web* (ACM, 2008), pp. 1265–1266.

[191] C. Bizer, T. Heath, D. Ayers and Y. Raimond. *Interlinking open data on the web*. In *Demonstrations Track, 4th European Semantic Web Conference, Innsbruck, Austria* (2007).

[192] S. A. Hart. *Apprentissage formel, informel, non-formel, des notions difficiles a utiliser... pourquoi ?* L'OBSERVATOIRE COMPETENCES-EMPLOIS sur la formation continue et le développement des compétences : http ://www.oce.uqam.ca/les-bulletins/90-notions-formel-informel-nonformel.html (BULLETIN JUIN 2013, volume 4, numéro 2 , consulté le 23 Juillet 2014).

[193] D. Poirier. *Des textes communautaires à la recommandation*. Ph.D. thesis, Université d'Orléans (2011).

[194] D. Brickley and L. Miller. *Foaf vocabulary specification 0.98*. Namespace Document 9 (2012).

[195] M. Lefevre and S. Jean-Daubias. *Intégration de données hétérogènes : un exemple pour la constitution de profils d'apprenants*. Intégration Technologique et Nouvelles Perspectives d'Usage p. 147 (2012).

[196] E. Dumbill. *Xml watch : Finding friends with xml and rdf*. IBM Developer Works, http ://www-106. ibm. com/developerworks/xml/library/x-foaf. html (2002).

[197] M. A. Hearst. *Automatic acquisition of hyponyms from large text corpora*. In *Proceedings of the 14th conference on Computational linguistics-Volume 2* (Association for Computational Linguistics, 1992), pp. 539–545.

[198] H. Suryanto and P. Compton. *Discovery of ontologies from knowledge bases*. In *Proceedings of the 1st international conference on Knowledge capture* (ACM, 2001), pp. 171–178.

[199] A. Delteil, C. Faron-Zucker and R. Dieng. *Learning ontologies from rdf annotations.* In *Workshop on Ontology Learning* (2001).

[200] G. Modica, A. Gal and H. M. Jamil. *The use of machine-generated ontologies in dynamic information seeking*. In *Cooperative Information Systems* (Springer, 2001), pp. 433–447.

[201] N. Stojanovic, L. Stojanovic and R. Volz. *A reverse engineering approach for migrating data-intensive web sites to the semantic web*. In *Intelligent Information Processing* (Springer, 2002), pp. 141–154.

[202] V. Kashyap. *Design and creation of ontologies for environmental information retrieval*. In *Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management* (Citeseer, 1999), pp. 1–18.

[203] J. Rebeyrolle. *Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes*. Actes Journées Francophones d'Ingénierie de la Connaissance IC'2000 pp. 105–114 (2000).

[204] A. Maedche and S. Staab. *The text-to-onto ontology learning environment*. In *Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures* (2000).

[205] A. Mikheev and S. Finch. *A workbench for finding structure in texts*. In *Proceedings of the fifth conference on Applied natural language processing* (Association for Computational Linguistics, 1997), pp. 372–379.

[206] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning. *Kea : Practical automatic keyphrase extraction*. In *Proceedings of the fourth ACM conference on Digital libraries* (ACM, 1999), pp. 254–255.

[207] F. Xu, D. Kurz, J. Piskorski and S. Schmeier. *A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping.* In *LREC* (2002).

[208] C. A. Thompson and R. J. Mooney. *Semantic lexicon acquisition for learning parsers*. Technical Note. January (1997).

[209] K. M. Gupta, D. W. Aha, E. Marsh and T. Maney. *An architecture for engineering sublanguage wordnets*. In *Proceedings of the First International Conference On Global WordNet* (2002), pp. 207–215.

[210] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning. *Kea : Practical automatic keyphrase extraction*. In *Proceedings of the fourth ACM conference on Digital libraries* (ACM, 1999), pp. 254–255.

[211] A. Gal, G. Modica and H. Jamil. *Ontobuilder : Fully automatic extraction and consolidation of ontologies from web sources*. In *Data Engineering, 2004. Proceedings. 20th International Conference on* (IEEE, 2004), p. 853.

[212] A. Oliveira, F. C. Pereira and A. Cardoso. *Automatic reading and learning from text*. In *Proceedings of the International Symposium on Artificial Intelligence (ISAI)* (Citeseer, 2001).

[213] E. Alfonseca and S. Manandhar. *An unsupervised method for general named entity recognition and automated concept discovery*. In *Proceedings of the 1st International Conference on General WordNet, Mysore, India* (2002), pp. 34–43.

[214] P. Velardi, R. Navigli, A. Cucchiarelli and F. Neri. *Evaluation of ontolearn, a methodology for automatic learning of domain ontologies*. Ontology Learning and Population (2005).

[215] B. Bachimont, A. Isaac and R. Troncy. *Semantic commitment for designing ontologies : a proposal*. In *Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web* (Springer, 2002), pp. 114–121.

[216] S.-H. Wu and W.-L. Hsu. *Soat : a semi-automatic domain ontology acquisition tool from chinese corpus*. In *Proceedings of the 19th international conference on Computational linguistics-Volume 2* (Association for Computational Linguistics, 2002), pp. 1–5.

[217] D. Faure and T. Poibeau. *First experiments of using semantic knowledge learned by asium for information extraction task using intex*. In *Proceedings of the ECAI2000 Ontology Learning Workshop* (Citeseer, 2000).

[218] G. Bisson, C. Nédellec and D. Canamero. *Designing clustering methods for ontology building-the mo'k workbench.* In *ECAI workshop on ontology learning* (Citeseer, 2000), volume 31.

[219] G. de Chaelandar and B. Grau. *Svetlan-a system to classify words in context.* In *ECAI Workshop on Ontology Learning* (2000).

[220] B. Biebow, S. Szulman and A. J. Clément. *Terminae : A linguistics-based tool for the building of a domain ontology*. In *Knowledge Acquisition, Modeling and Management* (Springer, 1999), pp. 49–66.

[221] J. Poyet. *Dimensions des représentations du concept de temps dans treize classes du préscolaire et du premier cycle du primaire au québec* (2010).

[222] D. Faure, C. Nédellec and C. Rouveirol. *Acquisition of semantic knowledge using machine learning methods : The system" asium"*. In *Universite Paris Sud* (Citeseer, 1998).

[223] L. Karoui, M.-A. Aufaure and N. Bennacer. *Extraction contextuelle de concepts ontologiques pour le web sémantique*. In *Actes de la conférence* (Citeseer, 2007).

[224] I. Bedini and B. Nguyen. *Automatic ontology generation : State of the art*. PRiSM Laboratory Technical Report. University of Versailles (2007).

[225] C. Romero, S. Ventura and E. Garcia. *Data mining in course management systems : Moodle case study and tutorial*. Computers & Education 51, 368 (2008).

[226] C.-N. Ziegler and G. Lausen. *Analyzing correlation between trust and user similarity in online communities*. In *Trust management* (Springer, 2004), pp. 251–265.