N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILALI LIABES
FACULTE DES SCIENCES EXACTES
SIDI BEL ABBÈS

THESE DE DOCTORAT DE 3ème CYCLE

Présentée par

FRIDI Asmaà

Domaine : Informatique
Filière : Système d'information et connaissances
Intitulé de la formation : Système d'information et connaissances

Intitulée

Approche sémantique pour le développement des systèmes de recommandation

Soutenue le……
Devant le jury composé de :

Président : Dr.GAFOUR Abdelkader, *MCA,* Université Djillali Liabès de Sidi-Bel-Abbès.
Examinateurs : Dr. TOUMOUH Adil, *MCA,* Université Djillali Liabès de Sidi-Bel-Abbès.
Dr. KESKES Nabil, *MCA,* École Supérieure en Informatique de Sidi Bel Abbes.
Dr. ADJOUDJ Réda, *MCA,* Université Djillali Liabès de Sidi-Bel-Abbès.
Directeur de thèse : : Pr. Sidi Mohamed BENSLIMANE, Pr, École Supérieure en Informatique de Sidi Bel Abbes.

Année universitaire *2016/2017*

Dédicace

À ma chère mère,

À mon père,

À mon frère,

Et á tous qui me sont chers.

Remerciement

Soyons reconnaissants aux personnes qui nous donnent,

du bonheur; elles sont les charmants jardiniers ,

par qui nos âmes sont fleuries

Marcel Proust

Le temps met tout en lumière.

Thalès

Le seul moyen de se délivrer d'une tentation, c'est d'y céder paraît-il ! Alors j'y cède en disant en grand Merci aux personnes qui ont cru en moi et qui m'ont permis d'arriver au bout de cette thèse.

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

J'exprime tous mes remerciements à l'ensemble des membres de mon jury pour leur disponibilité et acceptation de faire partie de ce jury, d'examiner et de rapporter mon travail.

Je remercie tous les membres du Laboratoire de Recherche en Informatique de Sidi Bel-Abbes (LABRI-SBA) .

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

Enfin, les mots les plus simples étant les plus forts, j'adresse toute mon affection à ma famille, et en particulier à ma maman qui m'a fait comprendre que la vie n'est pas faite que de problèmes qu'on pourrait résoudre grâce à des formules mathématiques et des algorithmes. Malgré mon éloignement depuis de (trop) nombreuses années, son intelligence, son confiance, son tendresse, son amour me portent et me guident tous les jours. Merci pour avoir fait de moi ce que je suis aujourd'hui. Est-ce un bon endroit pour dire ce genre de choses ? Je n'en connais en tous cas pas de mauvais. Je vous aime.

Résumé

Les systèmes de recommandations ont contribué à la réussite des sites Web personnalisés car ils peuvent automatiquement et efficacement choisir des articles ou des services adaptés à l'intérêt de l'utilisateur à partir d'énormes ensembles de données. Cependant, ces systèmes souffrent de problématiques liées au nombre peu important d'évaluations, au démarrage à froid du système et au nouvel utilisateur et nouvelle ressource. C'est pour cela que plusieurs voies ont été explorées pour trouver des solutions aux problématiques associées.

Le World Wide Web évolue d'un Web des documents hyper-liés à un Web des données liées. L'avènement de l'initiative Linked Open Data (LOD) a donné naissance à un grand choix de bases de connaissances ouvertes librement accessibles sur le Web tel que DBpedia. Ils fournissent une source d'information précieuse qui peut améliorer les systèmes de recommandation conventionnels, si elle est bien exploitée.

Dans cette thèse, nous montrons que l'utilisation des informations sémantiques décrivant les utilisateurs et les ressources peuvent améliorer la précision, la couverture et la qualité des systèmes de recommandation. Ceci en fournissant des informations supplémentaires permettant d'enrichir les items et les utilisateurs. Le web sémantique ou plus précisément le web des données ouvertes liées est l'infrastructure idéale pour obtenir ces descriptions sémantiques, car il permet la gestion décentralisée de l'information et l'agrégation de plusieurs sources.

Abstract

Recommendation systems have contributed to the success of custom websites as they can automatically and efficiently choose articles or services tailored to the user's interest from huge data sets. However, these systems suffer from problems related to the small number of evaluations, the cold start of the system and the new user, new resource. This is why several avenues have been explored to find solutions to the associated problems.

The World Wide Web moves from a web hyper-linked documents to a web-related data. The advent of the Linked Open Data (LOD) initiative gave rise to a wide array of open-source, open source web-based knowledge bases such as DBpedia. They provide a valuable source of information that can improve conventional referral systems, if properly exploited.

In this thesis, we show that the use of semantic information describing users and resources can improve the accuracy, coverage and quality of referral systems. This provides additional information to enrich items and users. The semantic web or more precisely the web of linked open data is the ideal infrastructure for obtaining these semantic descriptions because it allows the decentralized management of information and the aggregation of several sources.

Table des matières

Dédicace **2**

Remerciement **3**

Résumé **5**

Abstract **6**

Introduction générale **14**

1 Les systèmes de recommandation **19**
1.1 Introduction 19
1.2 Définition 19
1.3 Historique 20
1.4 Processus de recommandation 21
1.5 La classification des systèmes de recommandation 22
1.5.1 Recommandation démographique 23
1.5.2 Recommandations à base de connaissances 24
1.5.3 Recommandation communautaire 25
1.5.4 Recommandation basée sur le contenu 26
1.5.5 Recommandation collaborative (Filtrage collaboratif) . . . 30
1.5.6 Recommandation hybride 33
1.6 Forces et faiblesses des méthodes de recommandations 34
1.7 Architecture du système de recommandation 38
1.7.1 La collecte d'information 38
1.7.2 Modèle Utilisateur 38
1.7.3 Liste de recommandations 40
1.8 évaluation des systèmes de recommandation 40
1.9 Conclusion 42

2 Web sémantique et Linked data **43**
2.1 Introduction . . . 43
2.2 Historique . . . 44
2.3 Les Données ouvertes (open data) . . . 47
2.3.1 Pourquoi les Open Government Data ? . . . 48
2.3.2 Les Caractéristiques des données publiques ouvertes (Open Government Data) . . . 49
2.3.3 Les données cibles à l'ouverture . . . 51
2.3.4 Les cinq étoiles de l'open data . . . 51
2.4 Le web Sémantique . . . 52
2.4.1 Les couches du web sémantique . . . 53
2.4.2 Technologie de web sémantique . . . 63
2.5 Définition des données liées . . . 68
2.5.1 Les principes de données liées . . . 69
2.5.2 Nommer des éléments avec des URI . . . 70
2.5.3 Rendre les URI déréférençables . . . 71
2.6 Linked Open Data . . . 72
2.7 Les types de données . . . 72
2.8 Les bénéfices des données ouvertes liées . . . 72
2.9 Conclusion . . . 74

3 Les Systèmes de recommandation à base de LOD **75**
3.1 Introduction . . . 75
3.2 [Heitmann et al ,2010] . . . 75
3.2.1 Objectifs . . . 75
3.2.2 Principe de fonctionnement . . . 75
3.2.3 Points forts . . . 76
3.2.4 Points faibles . . . 77
3.3 [Ostuni et al, 2012] . . . 77
3.3.1 Objectifs . . . 77
3.3.2 Principe de fonctionnement . . . 77
3.3.3 Points forts . . . 79
3.3.4 Points faibles . . . 79
3.4 [Ostuni et al, 2013] . . . 79
3.4.1 Objectifs . . . 79
3.4.2 Principe de fonctionnement . . . 80
3.4.3 Les points forts . . . 81
3.4.4 Points faibles . . . 81
3.5 [Yang et al, 2013] . . . 81
3.5.1 Objectifs . . . 81
3.5.2 Le principe de fonctionnement . . . 82

3.5.3 Les points forts . . . 84
3.5.4 Les points faibles . . . 84
3.6 [PESKA et al,2013] . . . 85
3.6.1 Objectifs . . . 85
3.6.2 Le principe de fonctionnement . . . 85
3.6.3 Points forts . . . 86
3.6.4 Points faible . . . 86
3.7 [Ku et al, 2014] . . . 87
3.7.1 Objectifs . . . 87
3.7.2 Le principe de fonctionnement . . . 87
3.7.3 Points forts . . . 89
3.7.4 Points faibles . . . 89
3.8 [Ragone and al, 2017] . . . 90
3.8.1 Objectifs . . . 90
3.8.2 Le principe de fonctionnement . . . 90
3.9 Synthèse . . . 91
3.9.1 Filtrage Collaboratif . . . 91
3.9.2 Filtrage à base de contenu . . . 91
3.9.3 Filtrage Hybride . . . 92
3.10 Conclusion . . . 94

4 Approche sémantique pour l'amélioration des RS 95
4.1 Introduction . . . 95
4.2 Architecture générale du système . . . 95
4.3 Ressources . . . 96
4.3.1 Utilisateur . . . 97
4.3.2 Item . . . 97
4.3.3 Entité . . . 97
4.4 Principe de fonctionnement . . . 97
4.4.1 Préparation des données . . . 98
4.4.2 Recommandation des items . . . 104
4.5 Expérimentation . . . 107
4.5.1 Dataset . . . 107
4.5.2 Environnent d'expérimentation . . . 108
4.5.3 L'impact des méthodes de filtrage collaboratif sur le système de recommandation . . . 108
4.5.4 L'impact des méthodes de filtrage à base de contenu sur le système de recommandation . . . 109
4.5.5 Impact du nombre de cluster sur la recommandation . . . 111
4.5.6 Comparaison entre les différentes techniques de regroupement 111
4.6 Conclusion . . . 112

5 Conclusion et perspectives **113**
5.1 Conclusion 113
5.2 Perspectives 114

Bibliographie **115**

Table des figures

1.1 Processus de recommandation 22
1.2 Principales classifications des systèmes de recommandation 23
1.3 Recommandation démographique 24
1.4 Recommandation à base de connaissance 25
1.5 Recommandation Communautaire 26
1.6 Recommandation à base de contenu 27
1.7 Recommandation Collaborative 31
1.8 Exemple de recommandation Collaborative 32

2.1 Evolution du web 47
2.2 Diagramme de l'open government 49
2.3 Web sémantique 53
2.4 Architecture en couches du web sémantique [PLU, 2011]. 53
2.5 Exemple d'une représentation RDF 57
2.6 Exemple d'un graphe RDF 58
2.7 Exemple d'un graphe RDF avec des URI 58
2.8 URI sont utilisés pour identifier les personnes et les relations entre eux 71

3.1 Traitement de données liées pour les recommandations collaboratives 77
3.2 La représentation matricielle d'un graphe RDF de domaine de film. 78
3.3 Combinaison des graphes (étape 1 et 2) 82
3.4 Vue d'ensemble de l'approche 84
3.5 L'architecture de l'amélioration du système de recommandation avec LOD. 85
3.6 Processus global de l'approche 87
3.7 Processus global de l'approche 89

4.1 Architecture générale du Système 96
4.2 Le graphe de similarité des utilisateurs 100
4.3 La représentation matricielle d'un graphe RDF 102

4.4 Le modèle de donnée (graphe sémantique) 104
4.5 Exemple de Chemin collaboratif 106
4.6 Exemple de chemin à base de contenu 106
4.7 Exemple de Chemin hybride 106
4.8 Interface principale de notre système de recommandation 108
4.9 L'évaluation des méthodes de filtrage collaboratif 109
4.10 L'évaluation de l'hybridation du graphe d'informations personnelles et le graphe de classement 110
4.11 L'évaluation du VSM 110
4.12 Impact du nombre de cluster sur la recommandation 111
4.13 Comparaison entre les différentes techniques de regroupement . . . 112

Liste des tableaux

1.1 Les méthodes hybrides, adaptées de [Burke, 2002] 35
1.2 Les avantages et les inconvénients des méthodes de recommandation 37
1.3 Avantages et Inconvénients de la collecte explicite et implicite . . . 39
1.4 Les attributs qualitatifs pour évaluer la qualité d'un système de recommandation 41
1.5 Les attributs qualitatifs pour évaluer la qualité d'un système de recommandation 42

2.1 Extension du Web de Documents vers le Web Sémantique 63

3.1 Tableau comparatif entre les différentes approches de système de recommandation basé sur les LOD 93

Introduction générale

Contexte

Avec la propagation généralisée des données publiées dans le Web 2.0, nous sommes entrés dans une ère de surcharge d'information : plus l'information est produite plus nous ne pouvons pas vraiment la consommer et la traiter. Afin de faire face à un tel problème, il y a eu un nombre croissant de systèmes de filtrage qui tentent de soutenir leurs utilisateurs lors de la recherche d'informations. Parmi eux, nous mentionnons une famille spécifique de ces systèmes de filtrage de l'information : systèmes de recommandations, dont L'objectif est d'exploiter le maximum de données disponibles dans le Web pour répondre aux besoins des utilisateurs

Les systèmes de recommandation ont émergé comme un domaine de recherche à part entière au milieu des années 1990. Bien que le domaine de recherche des systèmes de recommandation soit plus récent et moins répandu que le filtrage d'information et la recherche documentaire et en particulier les moteurs de recherche, son intérêt dans l'industrie des e-services n'en est pas moins important. Les systèmes de recommandation ont été utilisés dans de nombreux domaines recommandant toutes natures d'items : pages Web, images, livres, news, films, musique, restaurants, objets, etc. Les systèmes de recommandation ont vu leur popularité croître ces dernières années en raison de la démocratisation du web et de l'augmentation exponentielle de la quantité de ressources disponibles et accessibles tels qu'Amazon, Netflix [Brun et al., 2014].

Les systèmes de recommandation ont été introduits comme une technique intelligente pour faire le filtrage de l'information afin de présenter les éléments d'information qui sont susceptibles d'intéresser l'utilisateur. Ils peuvent être utilisés pour fournir efficacement des services personnalisés dans la plupart des domaines de commerce électronique, en tant que client ou commerçant. Les systèmes de recommandation sont bénéfiques pour le client en lui faisant des suggestions sur les produits susceptibles d'être appréciés. En même temps, l'entreprise va bénéficier

de l'augmentation des ventes qui se produit normalement quand on présente au client plus d'articles susceptibles d'être aimés [Margaritis et al., 2003] .

Les systèmes de recommandation sont des systèmes capables de fournir des recommandations utiles ou intéressantes aux utilisateurs à partir de données multiples. Dans un processus de recommandation, l'identification des appréciations des utilisateurs est souvent fondamentale, dans la mesure où elle permet de connaitre l'utilisateur afin de lui proposer des recommandations pertinentes. Les appréciations reflètent les avis positifs ou négatifs des utilisateurs vis-à-vis d'un certain nombre d'items.

Problématique

Les données disponibles sur le Web en grandes quantités nécessitent un filtrage pour sélectionner les meilleures données, plusieurs algorithmes et approches ont été proposées dans La littérature, mais pas assez, ces approches souffrent de plusieurs problèmes tels que démarrage à froid, la sparcity, nouvel utilisateur, nouvel item, etc.

Le projet « Linking Open Data » commencé comme un effort de groupe en 2007, et a aidé à produire des milliards de déclarations RDF qui sont maintenant publiées sur le Web.

Le résultat de cette initiative est une énorme base de connaissance décentralisée, communément appelée le nuage LOD, dans lequel chaque «morceau de peu de connaissances» est enrichi par des liens vers des données relatives. Avec la plus grande disponibilité de cette connaissance, il y a un grand intérêt en tirant profit d'une telle information pour améliorer la qualité des systèmes de recommandations classiques. Bien que les systèmes de recommandation accroissent des technologies et des outils bien établis, de nouveaux défis se posent dans l'exploitation d'énorme quantité de données interconnectées provenant du Web de données. Dans le passé, plusieurs ouvrages sur les systèmes de recommandation ontologiques ont été proposés. En particulier, ils se sont avérés très efficaces en résolvant quelques inconvénients des méthodes de filtrages. Pour la création d'une nouvelle architecture de systèmes recommandations, il faut en premier lieu choisir quel type de système de recommandation il est préférable de réaliser : système basé sur le contenu, collaboratif ou hybride ? Ce choix dépend très fortement du niveau de qualité de l'information contenue dans le système d'information du fournisseur et des critères de recommandations des items entre les utilisateurs. Nous partons du postulat que le contenu de notre système de recommandation est modélisé à partir de données

sémantique, issues des données liés, et par conséquent, la qualité du contenu est avérée.

L'objectif principal de notre travail consiste à réaliser un système de recommandation visant à introduire l'aspect sémantique en recueillant des données auprès des LOD afin de minimiser les problèmes mentionnés ci-dessous.

Contribution

La couche sémantique est la partie qui permet de modéliser le contenu, la connaissance selon le domaine et selon les besoins utilisateurs dans l'application. Dans cette modélisation, l'apport des technologies du Web sémantique est significatif. En effet, l'utilisation des données liées pour sémantiser le contenu va permettre la modélisation de réseaux complexes de connaissances. D'autres parts, l'utilisation des informations sémantiques décrivant les utilisateurs et les ressources peut améliorer la précision, la couverture et la qualité des systèmes de recommandation. Ceci en fournissant des informations supplémentaires permettant d'enrichir les similarités collaboratives calculées à partir des évaluations par des similarités sémantiques.

Dans cette thèse, Nous proposons une nouvelle approche qui recueille des données de Linked Open Data (LOD) afin d'améliorer la précision et la qualité des systèmes de recommandation. Ceci est réalisé par un processus divisé en deux phases. La première phase comprend l'enrichissement des items à partir de DBpedia, le regroupement de ces items en se basant sur de leurs descriptions sémantiques et la génération de modèle de données dont chaque utilisateur et reliées par l'ensemble des items qui a déjà évalué où chaque arc est representé par le feedback donné par l'utilisateur à l'item dans un contexte donné.

La deuxième phase consiste à générer, filtrer et classer les chemins en se basant sur le modèle de donnée et le regroupement faites dans la première phase. Enfin, les principaux éléments ciblés par les meilleurs chemins sont recommandés à l'utilisateur. Nos contributions se résument aux points suivants :

Nous avons élaboré un état de l'art des différentes approches ont été proposées dans la littérature pour incorporer les techniques du Web sémantique au système de recommandation plus précisément les techniques de données ouvertes liées LOD. Un tableau comparatif faisant ressortir les différents points forts et faibles de chaque approche nous a permis de mieux positionner notre contribution. Nous

avons montré que l'inclusion des informations sémantiques, selon l'approche décrite dans notre proposition améliorait la précision et la couverture de l'algorithme de recommandation. Pour cela nous avons évalué plusieurs algorithmes : collaboratif pur, sémantique pur ainsi que plusieurs stratégies d'hybridation sur un jeu de données que nous avons enrichi d'informations supplémentaires extraites à partir de sources sémantiques.

Pour la validation de l'approche proposée dans cette thèse, nous avons adopté une démarche empirique consistant à effectuer une série d'expérimentations sur des données réelles issues du jeu de données Movielens (corpus de référence largement exploité par la communauté scientifique travaillant sur les systèmes de recommandation). Nous avons couplé ces données avec les données issues de la base de données DBpedia pour extraire les informations sémantiques sur les items. Le choix du corpus Movielens, nous a permis de confronter certains de nos résultats avec ceux de la communauté scientifique. Les résultats obtenus ont été évalués en termes de précision et de qualité de recommandation.

Organisation

En plus d'une introduction générale et une conclusion, cette thèse comporte quatre chapitres qui se répartissent comme suit :

- Le premier chapitre présente les systèmes de recommandation. Il introduit l'histoire des systèmes de recommandation, suivie de la présentation de plusieurs classifications bien connues dans ce domaine.

- Le deuxieme chapitre est composé de quatre parties. La première partie est consacrée aux données ouvertes où nous allons citer les diffèrent caractéristiques de ces données, les données de gouvernement. La deuxième partie traitera le web sémantique et ses différentes couches et différentes technologies. La troisième partie, concerne les données liées et ces principes. La dernière partie, traitera les données ouvertes liées, ses types en précisant les bénéfices de ces derniers.

- Dans le chapitre trois, nous dressons un état de l'art des différentes approches relatives à l'amélioration des systèmes de recommandations fondés sur les données ouvertes liées. Un tableau comparatif nous a permis de mieux positionner nos contributions.

- Le quatrième chapitre introduit l'approche proposée durant la thèse. En premier lieu l'architecture globale est présentée. Ensuite, les différentes phases de

l'approche son explicitées. Enfin, les résultats obtenus sont illustrés, analysés, et comparés avec ceux des approches connexes

Chapitre 1

Les systèmes de recommandation

1.1 Introduction

Dans tous les temps les sociétés humaines ont produit, stocké et échangé des données, des informations dans le but de transmettre des connaissances. Aujourd'hui, ces données sont majoritairement numériques, et elles sont stockées sur des ordinateurs et échangées sur des réseaux de télécommunication informatiques comme Internet mais la quantité d'informations hétérogènes stockées de manière numérique double tous les dix-huit mois citant par exemple le réseau sociale Facebook qui draine plus de 1,8 milliard de comptes internet dans le monde

Cette croissance exponentielle des données se traduit par une difficulté à organiser et à analyser ces informations brutes ouvrant pourtant de nouvelles voies sur les chemins de la connaissance. La question n'est donc plus de disposer de l'information, mais de trouver l'information pertinente au bon moment. Delà, le système de recommandation est né dont sa qualité est étroitement liée à sa capacité à prendre en compte et à traiter une grande quantité d'évaluation, qu'est-ce qu'un système de recommandation ? et comment peut-il satisfaire les besoins et les préférences des usagers ?

1.2 Définition

Les systèmes de recommandation peuvent être définis de plusieurs façons, vue la diversité des classifications proposées pour ces systèmes, mais il existe une définition générale de Robin Burke [Burke, 2002] qui les définit comme suit :

« Des systèmes capable de fournir des recommandations personnalisées permettant de guider l'utilisateur vers des ressources intéressantes et utiles au sein d'un espace de données important ».

Autrement dit, c'est une manière de proposer à l'utilisateur des produits qui sont susceptibles de l'intéresser.

Les deux entités de base qui apparaissent dans tous les systèmes de recommandations sont l'usager et l'item. L' «usager» est la personne qui utilise un système de recommandation, donne son opinion sur diverses items et reçoit les nouvelles recommandations du système. L'«Item» est le terme général utilisé pour désigner ce que le système recommande aux usagers.

Les données d'entrée pour un système de recommandation dépendent du type de l'algorithme de filtrage employé. Généralement, elles appartiennent à l'une des catégories suivantes

- **Les estimations :**(également appelées les votes), expriment l'opinion des utilisateurs sur les articles (exemple : 1 mauvais à 5 excellent).
- **Les données démographiques :** se réfèrent à des informations telles que l'âge, le sexe, le pays et l'éducation des utilisateurs. Ce type de données est généralement difficile à obtenir et est normalement collecté explicitement ;
- **Les données de contenu :** qui sont fondées sur une analyse textuelle des documents liés aux éléments évalués par l'utilisateur. Les caractéristiques extraites de cette analyse sont utilisées comme entrées dans l'algorithme de filtrage afin d'en déduire un profil d'utilisateur [Margaritis and al., 2003].

1.3 Historique

Les racines des systèmes de recommandation remontent aux travaux étendus dans les sciences cognitives, la théorie d'approximation, la recherche d'informations, la théorie de la prévoyance et ont également des liens avec la science de la gestion et le marketing.

« Information Lens System » [Malone et al., 1987] peut être considéré comme le premier système de recommandation. À l'époque, l'approche la plus commune pour le problème de partage d'informations dans l'environnement de messagerie

électronique était la liste de distributions basée sur les groupes d'intérêt. La première définition pour le filtrage a été donnée aussi par Malone :

« Même si le terme a une connotation littérale de laisser les choses dehors (filtrage négatif : enlèvement), nous l'utilisons ici dans un sens plus général qui consiste à sélectionner les choses à partir d'un ensemble plus large de possibilités (filtrage positif : sélection) ».

En 1992, La littérature académique a introduit le terme de filtrage collaboratif par le système « Tapestry » [Goldberg et al, 1992], qui a permis aux utilisateurs de créer des requêtes permanentes, basées sur les annotations des utilisateurs. Quelques années plus tard, un certain nombre de systèmes académiques de recommandation ont été introduits dans les domaines de la musique [Maes et al, 1995], des livres [Resnick and al, 1997], des vidéos [Furnas and al., 1995], des films, des pages Web [Das and al., 1997], des articles de nouvelles Usenet [Konstan and al., 1997] et des liens Internet [Terveen and al., 1997].

Ainsi, depuis l'adoption d'Amazon, la technologie de recommandation, souvent basée sur le filtrage collaboratif, a été intégrée dans de nombreux systèmes de commerce électronique en ligne. Une motivation significative pour atteindre cet objectif est d'augmenter le volume de ventes, le client peut acheter un article si on le lui suggère, mais ne pourrait pas le rechercher autrement. Plusieurs entreprises, comme NetPerceptions et Strands, ont été construites pour fournir une technologie et des services de recommandation aux détaillants en ligne.

La boîte à outils de techniques de recommandation s'est également développée au-delà du filtrage de collaboration pour inclure des approches basées sur contenu, basées sur l'inférence de recherche d'informations, bayésienne et les méthodes basées sur le cas de raisonnement. Les systèmes hybrides de recommandation ont également émergé pendant que les diverses stratégies de recommandation ont mûri.

1.4 Processus de recommandation

En général, chaque système de recommandation fait suite à un processus spécifique pour produire des recommandations, voir la Figure 2.2. Les approches de recommandation peuvent être classées sur la base des sources d'informations qu'ils utilisent. Trois sources d'informations peuvent être identifiées comme entrée pour le processus de recommandation. Les sources disponibles sont les données d'utilisateur (données démographiques), les données de l'élément (mots-clés, genres) et

les évaluations utilisateur-item (obtenues par des données de transactions, les cotes explicites et implicites). Ces sources seront discutées ci-dessous [Anderson, 2011].

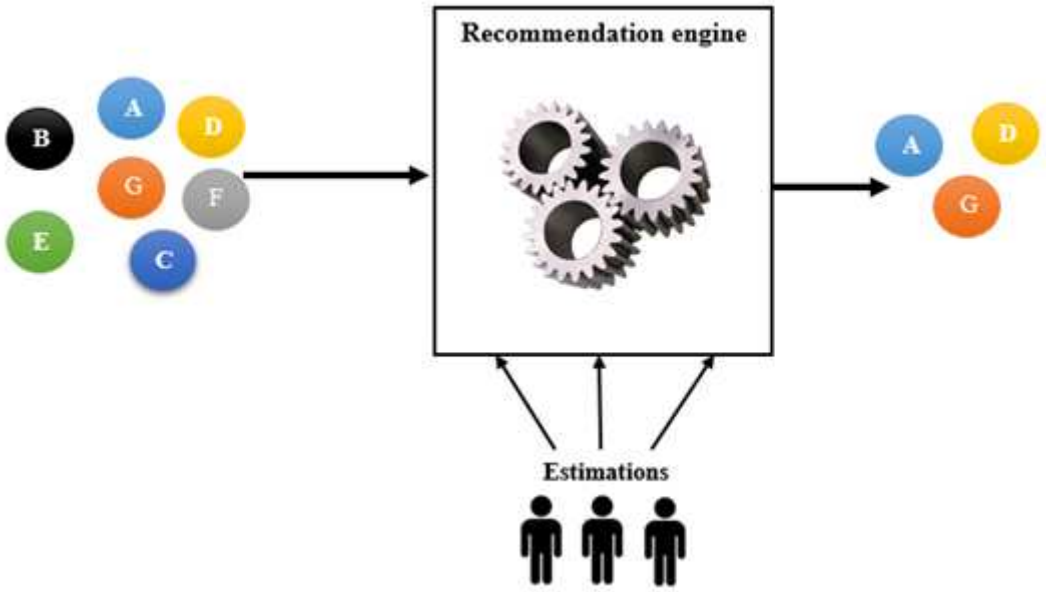

FIGURE 1.1 – Processus de recommandation

1.5 La classification des systèmes de recommandation

Dans la littérature plusieurs classifications des systèmes de recommandations sont connus (voir la Figure 1.2) :

1. **La classification classique** : cette classification de [Adomavicius et al, 2005] est identifiée par un filtrage collaboratif(CF), un filtrage basé sur le contenu(CBF) et le un filtrage hybride.

2. **La Classification de [Su et al, 2009]** :c'est une classification utilisée dans les systèmes purement collaboration. Ils proposent une sous-classification qui comprend les techniques hybrides les classer dans les méthodes de collaboration hybrides. [Su et al, 2009] classent FC en trois catégories :

   - ***Approches CF à base de mémoire*** : pour K-plus proches voisins.
   - ***Approches FC basé sur un modèle*** englobant une variété de techniques telles que : clustering, les réseaux bayésiens, factorisation de matrices, les processus de décision de Markov.

- ***CF hybride*** qui combine une technique recommandation CF avec un ou plusieurs autres méthodes.

3. **La classification de [Rao et al, 2008]** : c'est une classification en fonction de la source d'information utilisée.

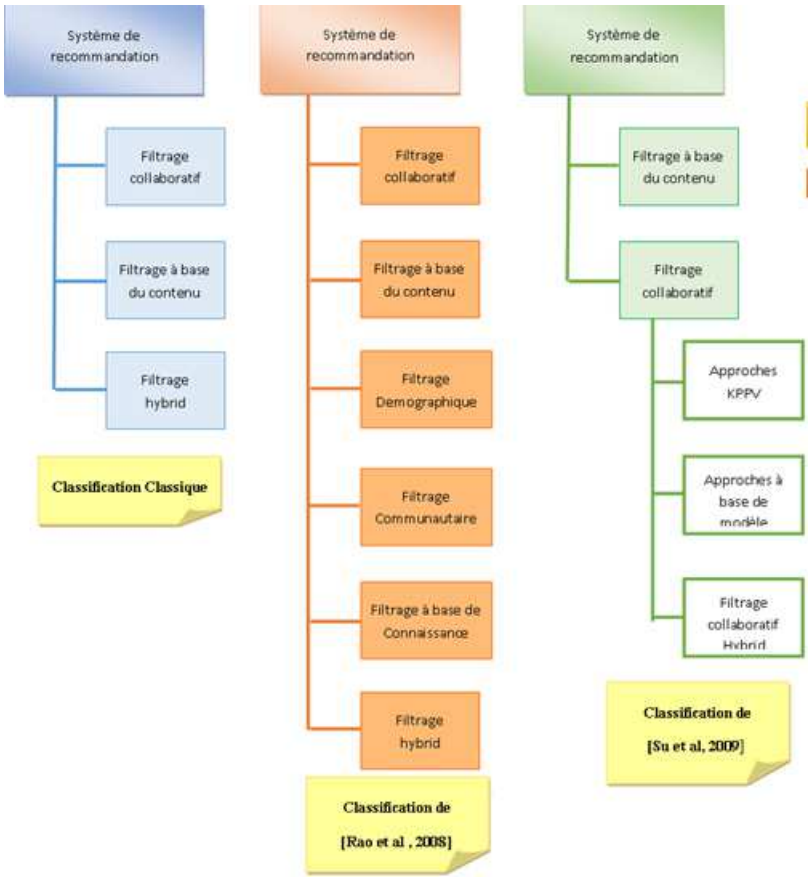

FIGURE 1.2 – Principales classifications des systèmes de recommandation

1.5.1 Recommandation démographique

Elle s'agit d'une recommandation simple qui propose des items par rapport au profil démographique d'usager (Figure 1.4). Elle consiste à répartir les usagers en plusieurs classes (groupes) en fonction d'informations démographiques telles que le sexe, l'âge, la profession, la localisation, la langue, le pays, etc. L'hypothèse sur laquelle repose cette approche est que deux utilisateurs ayant évolué dans un environnement similaire partagent des goûts communs que deux utilisateurs ayant

évolué dans des environnements différents et ne partageant donc pas les mêmes codes [Rochlitz, 1992]. De nombreux sites utilisent cette solution simple à proposer une offre de contenu "personnalisé". Par exemple, les utilisateurs sont redirigés vers un site Web particulier en fonction de leur langue ou de pays. Ces approches ont été très populaires dans la littérature de marketing, mais ont reçu peu d'attention dans le domaine des algorithmes de recommandation.

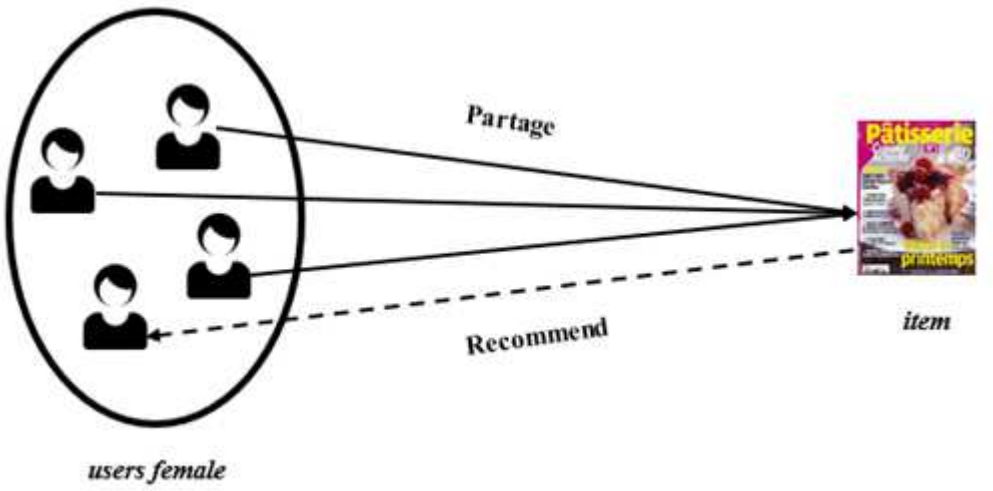

FIGURE 1.3 – Recommandation démographique

1.5.2 Recommandations à base de connaissances

Les systèmes à base de connaissances génèrent les recommandations en utilisant des connaissances spécifiques dont certaines caractéristiques d'items répondent aux préférences de l'usager (Figure 1.4).

Les systèmes à base de connaissances travaillent généralement mieux que les autres types de recommandation si les données limitées sont disponibles, à savoir, si le système ne peut pas compter sur l'existence d'un historique de l'utilisateur. Mais si le système de la connaissance n'est pas conçu pour apprendre des notes ou des actions de l'utilisateur.

Plusieurs approches ont été utilisées dans cette topologie de recommandations telles que le raisonnement à base des cas et le raisonnement à base de contraintes.

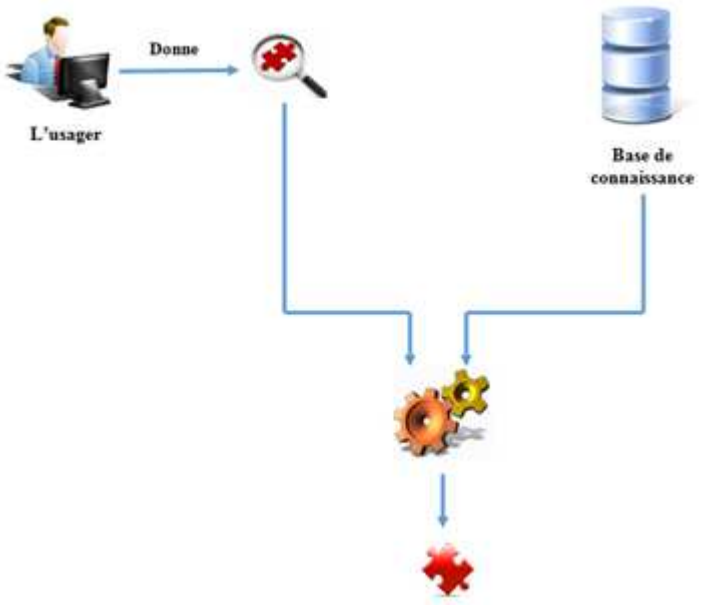

FIGURE 1.4 – Recommandation à base de connaissance

1.5.2.1 Le raisonnement à base des cas

Le raisonnement à base des cas tire parti de la régularité du monde réel afin de résoudre des problèmes en recherchant la solution d'un cas semblable rencontré et résolu dans le passé. [Piamrat et al, 2009] ont utilisé cette approche dans les systèmes de recommandation, ils estiment combien les besoins ou les préférences (description de problème) de l'usager correspondent aux recommandations possibles (solutions du problème) en se basant sur le comportement de consommation précédente (cas précédents).

1.5.2.2 Le raisonnement à base de contraintes

Une recommandation à base de contraintes est un autre type de systèmes à base de connaissances. La recommandation à base de contraintes exploite des bases de connaissances prédéfinies qui contiennent des règles explicites sur la façon de relier les exigences des clients avec des fonctionnalités d'item. Par exemple, un utilisateur peut être intéressé à acheter des produits avec un certain ensemble de caractéristiques et dans une gamme de prix spécifique.

1.5.3 Recommandation communautaire

La recommandation communautaire où comme on l'appelle souvent recommandation sociale vue que la plupart des réseaux sociaux (Facebook, Twitter, etc..) se

basent sur cette classification dans leurs recommandations.

Son principe général est que le système propose des recommandations à partir des relations de l'usager avec ces amis dans le réseau social, et parfois cette recommandation dépend aussi de la valeur de confiance d'usager dans chacun de ses amis, l'exemple le plus connu de cette recommandation est la section des pages et des groupes qui apparait dans la partie droite d'une page Facebook (Figure 1.5). L'importance décisionnelle du bouton « I Like » de Facebook a donné un succès croissant dont 55% des utilisateurs sont influencés par leurs amis. Facebook n'est

FIGURE 1.5 – Recommandation Communautaire

pas la seule illustration du développement de la recommandation sociale comme facteur de décision à l'achat de produits et de services. Nomao est un moteur de recherche géo localisée qui tient compte des recommandations des internautes dans l'affichage de ses résultats (tels que des commerces de proximité, restaurants, etc. . .). D'autres moteurs reposent sur des principes équivalents : comme Tumbup, TallSteet, GetGlue, ou encore Stumbleupon.

1.5.4 Recommandation basée sur le contenu

Le concept des systèmes de recommandation est d'orienter les usagers d'une manière personnalisée vers des objets intéressants issus d'un large espace d'options possibles. Les systèmes de recommandation basés sur le contenu proposent des items similaires à ceux aimés par l'utilisateur dans le passé. En effet, le processus principal réalisé par un système de recommandation basé sur le contenu consiste à faire correspondre les attributs d'un profil utilisateur dans lequel les préférences et

intérêts sont stockés avec les attributs des items tels que le titre, une description, des mots clés, la catégorie, etc. et qui peuvent être désignés comme métadonnées dans le but de recommander à l'utilisateur de nouveaux objets intéressants (Figure 1.6). Ce type de recommandation prend place dans différents domaine de recherche en

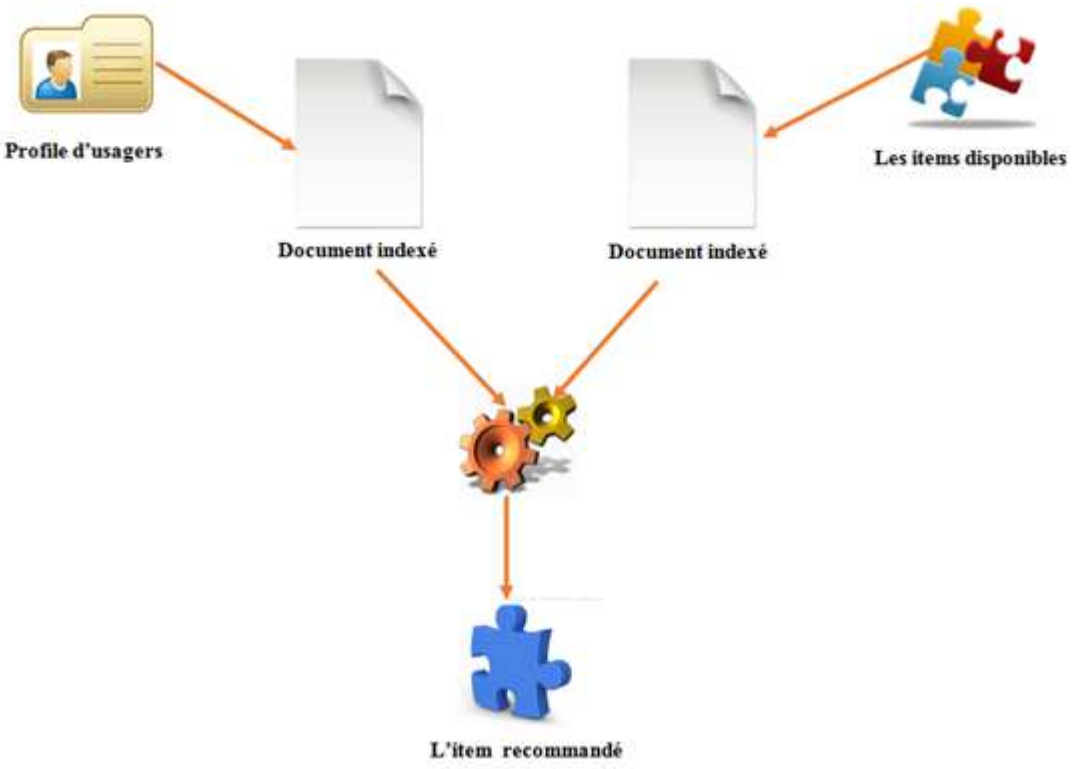

FIGURE 1.6 – Recommandation à base de contenu

informatiques tel que la Recherche d'Information (RI) et l'Intelligence Artificielle [Yates et al, 1999].

En Recherche d'Information, il est considéré que les utilisateurs voulant des recommandations sont engagés dans un processus de recherche d'information.

L'utilisateur exprime un besoin ponctuel en donnant une requête (habituellement une liste de mots-clefs), les items à recommander peuvent être très différents, en fonction du nombre et du type des attributs utilisés pour les décrire.

En intelligence artificielle, la tâche de recommandation peut être exprimée comme un problème d'apprentissage qui exploite la connaissance passée des utilisateurs.

De manière simple, les profils des utilisateurs sont sous la forme de vecteurs de mots-clefs et reflètent les intérêts à long terme de l'utilisateur.

Dans la plupart des systèmes de filtrage basé sur le contenu ; les items sont représentés par un ensemble de caractéristiques appelées aussi attributs ou propriétés par exemple les caractéristiques un film peuvent être : le genre, le sujet, les acteurs, etc. Ce sont des descriptions extraites à partir de pages web, des emails, des articles de news, ou des descriptions de produits. À partir de la représentation des items on peut distinguer deux types de recommandation basé sur le contenu : recommandation basé sur les mots clefs et recommandation basé sur la sémantique

1.5.4.1 Recommandation basée sur les mots clefs

La plupart des systèmes de recommandation utilisent de simples modèles de recherche, comme la correspondance de mots-clefs, ces systèmes ont été développés en très peu de temps dans de multiples domaines d'applications, comme les actualités, la musique, les films, etc. Chaque domaine présente différents problèmes qui requièrent différentes solutions.

Parmi les systèmes populaires qui utilisent ce type de recommandation, nous pouvons citer, **Letizia** [Liberman, 1995], un système implémenté comme une extension de navigateur Web traquant le comportement de l'usager et construit un modèle personnalisé constitué des mots-clefs liés aux intérêts de l'utilisateur.

Personal Web Watcher [Mladenic, 1999] apprend les intérêts des utilisateurs à partir des pages Web qu'ils visitent, et à partir des documents qui ont un lien hypermédia avec les pages visitées.

WebMate [Chen and Sycara, 1998] utilise une approche différente pour la représentation des intérêts de l'utilisateur. WebMate assure le suivi des intérêts de l'utilisateur dans différents domaines en construisant le profil de l'utilisateur par apprentissage. Ce profil est constitué de vecteurs de mot-clef qui représentent des exemples positifs d'apprentissage. Un profil de vecteurs de mot-clef peut représenter correctement jusqu'à intérêts indépendants des utilisateurs.

Your News [Ahn et al 2007], un système plus récent pour l'accès personnalisé aux actualités, garde un profil d'intérêt séparé pour 8 sujets différents (National, Monde, Business, etc.). Le profil d'intérêt de l'utilisateur pour chaque sujet est représenté avec un vecteur de termes prototypes pondérés, extraits de l'historique des actualités vues par l'utilisateur. Les articles des dernières actualités visionnées

par l'utilisateur sont collectés, et les 100 termes les plus pondérés sont extraits pour générer les vecteurs prototypes finaux. Le système considère des profils à court terme, en ne considérant que les 20 dernières actualités, alors que des profils à long terme utilisent tout ce qui a été vu. De l'analyse des principaux systèmes développés ces quinze dernières années, nous retenons que la représentation par mots-clefs à la fois pour les items et pour les profils peut donner des résultats précis.

La plupart des systèmes basés sur le contenu sont conçus comme des classificateurs de textes construits à partir d'un ensemble de documents d'apprentissage qui sont soit des exemples positifs, soit des exemples négatifs des intérêts de l'utilisateur. Le problème avec cette approche est le « manque d'intelligence ». Lorsque des caractéristiques plus complexes sont nécessaires, les approches à base de mots clefs montrent leurs limites. Si l'utilisateur, par exemple, aime « l'impressionnisme français », les approches à base de mots-clefs chercheront seulement des documents dans lesquels les mots « français » et « impressionnisme » apparaissent. Des documents concernant Claude Monet ou Renoir n'apparaîtront pas dans l'ensemble des recommandations, même s'ils sont susceptibles d'être pertinents pour l'utilisateur. Des stratégies de représentation plus avancées sont nécessaires pour que les systèmes de recommandation basés sur le contenu prennent en compte la sémantique associée aux mots.

1.5.4.2 Recommandation basée sur la sémantique

Plusieurs stratégies ont été utilisées pour introduire de la sémantique dans le processus de recommandation. La description de ces stratégies est abordée en tenant compte de plusieurs critères tels que :

- le type de source de connaissance impliquée (lexique, ontologie, etc.) ;
- les techniques adoptées pour l'annotation ou la représentation d'items ;
- le type de contenu inclus dans le profil utilisateur ;
- la stratégie de correspondance entre items et profil.

Les systèmes de recommandation basés sur la sémantique évoluent au rythme des méthodes et outils proposés dans le domaine du Web sémantique. **SiteIF** [Magnini et al, 2001] a été le premier système à adopter une représentation basée sur le sens des documents pour construire un modèle des intérêts de l'utilisateur SiteIF est un agent personnel pour un site Web de nouvelles multilingues. La source externe de connaissance impliquée dans le processus de représentation est MultiWordNet(une base de données lexicale multilingue)

ITR (ITermRecommender) est un système capable de fournir des recommandations d'items dans plusieurs domaines (films, musique, livres), à condition que les descriptions d'articles soient disponibles sous forme de documents texte [Degemmi et al, 2007]

SEWeP (Semantic Enhancement for Web Personalization) [Eirinak et al, 2003] est un système de personnalisation Web qui utilise à la fois les logs d'utilisation et la sémantique du contenu du site Web dans le but de le personnaliser. Une taxonomie des catégories spécifiques au domaine a été utilisée pour annoter sémantiquement les pages Web, afin d'avoir un vocabulaire uniforme et consistant

Quickstep[Middleton et al, 2004] est un système de recommandation d'articles de recherche académique. Le système adopte une ontologie d'articles de recherche basée sur la classification scientifique du projet DMOZ open directory (DMOZ open directory project) (27 classes utilisées).

InformedRecommender [Aciar, et al, 2007] utilise les avis des utilisateurs sur les produits pour faire des recommandations. Le système convertit les opinions des clients dans une forme structurée en utilisant une ontologie de traduction, qui est exploitée pour la représentation et le partage de connaissance.

Les méthodes décrites précédemment ont donné des résultats meilleurs et plus précis comparés aux méthodes traditionnelles basées sur le contenu

1.5.5 Recommandation collaborative (Filtrage collaboratif)

Contrairement à la recommandation basée sur le contenu, le filtrage collaboratif se base sur la construction profil d'usager qui sera créé au fur et à mesure de l'évaluation de l'usager, il permet de filtrer n'importe quel type de données (image, texte, vidéo). Le but du filtrage collaboratif est de suggérer de nouveaux items ou de prédire l'utilité d'items inconnus pour un utilisateur donné, en se fondant sur les évaluations déjà exprimées. Il utilise des méthodes statistiques pour faire des prévisions en se basant sur la corrélation entre son profil personnel et les profils des autres utilisateurs qui présentent des intérêts semblables (Figure 1.7). Un système de filtrage collaboratif est organisé comme suit :

- Collecter les appréciations et le comportement des utilisateurs, en général l'utilisateur fournit des évaluations sous forme de notes, sur un ou plusieurs axes : qualité, correspondance au besoin, etc. ;
- Intégrer ces informations au profil de l'utilisateur ;

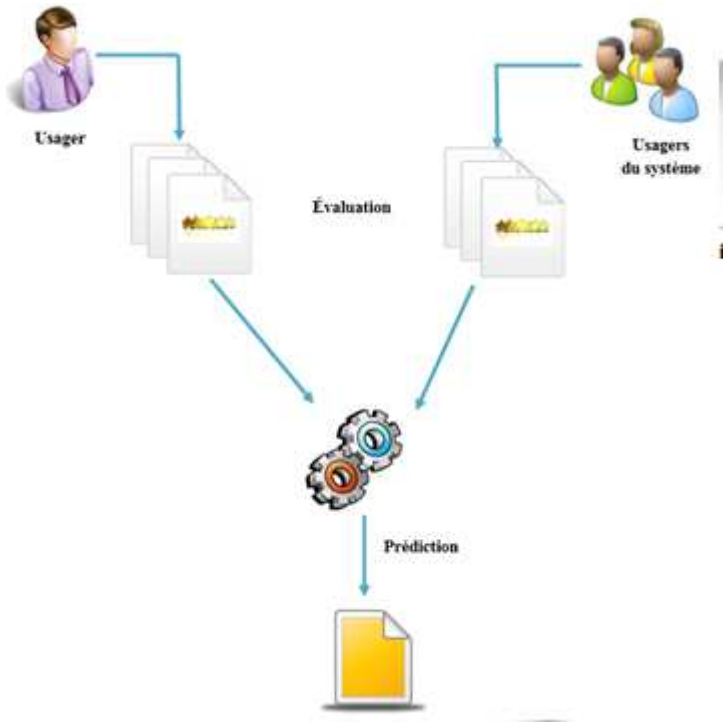

FIGURE 1.7 – Recommandation Collaborative

Le système utilise ces informations pour faire des recommandations.

La Figure 1.8 montre un tableau de films avec sur un axe les utilisateurs d'une même Système (ex : un groupe d'amis sur MovieLens) et sur un autre les films. Chaque cellule de la matrice contient l'avis donné par un utilisateur pour un film, la cellule vide signifie qu'il n'a pas d'avis particulier sur ce film. Pour pouvoir prédire si Illyes apprécierait le film "Harry Potter" et éventuellement lui recommander ce film, on compare les Votes d' Illyes à ceux des autres utilisateurs sélectionnés. On peut alors voir que Illyes et Amel ont des Votes identiques, et que Amel n'a pas aimé le film «Harry Potter », on pourrait alors prédire que Illyes n'aimera pas aussi ce film et de ne lui pas faire cette suggestion. Plusieurs systèmes utilisent le filtrage collaboratif vu les avantages qu'il apporte, parmi ces systèmes on peut citer : Amazon, Netflix, MovieLens, Jester, Citeseer, Tapestry, Phoaks, etc.

Plusieurs algorithmes sont utilisés dans ce type de recommandation,[Breese et al , 1998], les classés dans deux familles : les algorithmes basés mémoire est les algorithmes basés modèle mais on peut les combiner dans un algorithme hybride.

FIGURE 1.8 – Exemple de recommandation Collaborative

1.5.5.1 Les algorithmes à base de mémoire

Appelé aussi basé heuristique à base d'usager, utilise la totalité d'informations (évaluations des utilisateurs) disponible pour faire des prédictions. Ces prédictions sont faites par rapport aux usagers les plus proches voisins(PPV), pour choisir ces PPV ont doit calculer la similarité entre les usages en utilisant généralement deux mesures qui sont : la similarité vectorielle et la corrélation de Pearson Il existe aussi des algorithmes basés mémoire mais basé sur item où il calcule la similarité entre les items, l'item que l'usager à évaluer et l'item cible.

Bien que ce type d'algorithmes est sont très simple à mettre en œuvre et donnes des prédictions de bonne qualités, la forte complexité combinatoire les empêche d'être utilisés dans les environnements de production

1.5.5.2 Les algorithmes à base de modèle

Afin palier au problème de la complexité des algorithmes à base de mémoire, [Breese et al.,1998] ont proposés des approches à base de modèles où L'idée générale

est d'obtenir un modèle des données hors ligne pour prédire des estimations en ligne aussi vite que possible.

Pour la construction de ce modèle plusieurs méthodes ont été proposées en se basant sur les techniques d'apprentissage automatique et les techniques statistiques tels que le Clustering, réseaux des bayes, etc

1.5.6 Recommandation hybride

Afin de limiter les inconvénients (voir tableau 1.2) des différents types de recommandation cités ci-dessus, plusieurs approches qui ont été proposés, ce sont généralement une des combinaisons des deux classifications : recommandation à base du contenu, recommandation collaborative. Selon [Adomavicius et al, 2005] on peut distinguer quatre façons de combiner les méthodes traditionnelles :

1. Implémenter la méthode collaborative et la méthode basée sur le contenu séparément puis combiner leurs prédictions.
2. Incorporer quelques caractéristiques de la méthode basée sur le contenu dans l'approche collaborative.
3. Incorporer quelques caractéristiques de la méthode collaborative dans l'approche à base de contenu.
4. Construction d'un modèle général unifié qui incorpore les caractéristiques des deux modèles. [Naak, 2009].

La meilleure description des méthodes hybrides a été faite par [BURKE, 2002] (voir Tableau 1.1). Alors, selon Burke on peut distinguer sept façons de combiner les méthodes traditionnelles :

- **Pondération (Weighted)**
  Une méthode hybride qui combine la sortie d'approches distinctes, utilisant, par exemple, une combinaison linéaire des scores de chaque technique de recommandation.
- **Commutation (Switching)**
  C'est une technique qui permet de faire le choix d'un modèle de recommandation parmi plusieurs, en se basant sur plusieurs critères. La détermination de la technique appropriée dépend de la situation. Le système se doit alors de définir les critères de commutation, ou les cas où l'utilisation d'une autre technique est recommandée. Ceci permet au système de connaître les points forts et les points faibles des techniques de recommandation qui le constituent.

- **Technique mixte (Mixed)**
  Dans cette approche, le recommandeur ne combine pas, mais augmente la description des ensembles de données, en prenant en considération les estimations des utilisateurs et la description des articles. La nouvelle fonction de prédiction doit faire face aux deux types de descriptions et permet d'éviter les problèmes posés par le filtrage collaboratif.

- **Combinaison de caractéristiques (Features combination)**
  Dans un hybride basé sur la combinaison de caractéristiques, les données provenant de techniques collaboratives sont traitées comme une caractéristique, et une approche basée sur le contenu est utilisé sur ces données

- **Cascade**
  La cascade implique un processus étape par étape. Dans ce cas, une technique de recommandation est appliquée en premier, produisant un ensemble de candidats potentiels. Puis, une deuxième technique raffine les résultats obtenus dans la première étape. Cette méthode a pour avantage que si la première technique génère peu de recommandations, ou si ces recommandations sont ordonnées afin de permettre une sélection rapide, la deuxième technique ne sera plus utilisée.

- **Augmentation de caractéristiques (Feature augmentation)**
  L'augmentation de caractéristiques est semblable à la cascade, mais dans ce cas-là les résultats obtenus (le classement ou la classification) de la première technique sont utilisées par le deuxième comme une caractéristique ajoutée.

- **Méta niveau (Meta-level)**
  Dans un hybride basé sur méta niveau, une première technique est utilisée, mais différemment que la précédente méthode (augmentation de caractéristiques), non pas pour produire de nouvelles caractéristiques, mais pour produire un modèle. Et dans la deuxième étape, c'est le modèle entier qui servira d'entrée pour la deuxième technique.

1.6 Forces et faiblesses des méthodes de recommandations

Le tableau 1.2 résume les forces et faiblesses des méthodes traditionnelles, en l'occurrence le Filtrage Collaboratif (FC), le Filtrage Démographique(FD), le Filtrage à Base de Contenu (FBC), et le Filtrage à base de connaissance, Filtrage à base de données communautaires,

Technique	Description
Pondération	Les résultats (ou votes) des différents techniques de recommandation sont combinés pour produire une seule recommandation
Commutation	Le système commute entre les techniques de recommandation selon la situation actuelle
Technique mixte	Les recommandations de différents recommandeurs sont présentées en même temps
Combinaison des caractéristiques	Les données provenant d'une technique sont traitées comme une caractéristique, et une approche basée sur une autre technique est utilisée sur ces données
Cascade	Un recommandeur raffine les recommandations données par un autre
Augmentation des caractéristiques	La sortie d'une technique est utilisée comme une caractéristique d'entré à l'autre
Méta niveau	Le modèle appris d'un recommandeur est employé comme entrée à l'autre

TABLE 1.1 – Les méthodes hybrides, adaptées de [Burke, 2002]

- **Adaptabilité** : Au fur et à mesure que la base de données des évaluations augmente, la recommandation devient plus précise.
- **Nouvel usager** : un nouvel usager qui n'a pas encore accumulé suffisamment d'évaluations ne peut pas avoir de recommandations pertinentes.
- **nouvel item** : un item doit avoir suffisamment d'évaluations pour qu'il soit pris en considération dans le processus de recommandation.
- **Démarrage à froid** : le démarrage à froid est un problème pour les nouveaux usagers qui commencent à jouer avec le système, parce que le système ne dispose pas d'assez d'informations à leur sujet. Si le profil d'usager est vide, il doit consacrer une somme d'efforts à l'aide du système avant d'obtenir une récompense (les recommandations utiles). D'autre part, quand un nouvel élément est ajouté à la collection, le système doit avoir suffisamment d'informations pour être en mesure de recommander cet article aux utilisateurs.

La classification	Avantages	Inconvénients
Recommandation démographique	Un nouvel utilisateur peut avoir une recommandation sans n'avoir noté aucun item	• Problème de confidentialité • Recommandation très générale • Utilisateur avec un goût unique peut pas obtenu une bonne recommandation • Nouvel Item
Recommandation base de connaissances	*Adaptabilité* : Plus la base de connaissance est grande plus la recommandation est bonne	• Nouvel usager • Démarrage à froid • Nouvel Item
Recommandation communautaire	*Adaptabilité* : la qualité croit avec le nombre d'amis	• Nouvel usager • Nouvel Item
Recommandation à base du contenu	• pas besoin d'une large communauté d'utilisateurs pour pouvoir effectuer des recommandations • ne liste de recommandations peut être générée même s'il n'y a qu'un seul utilisateur.	• Nouvel usager • Nouvel item • L'analyse du contenu est nécessaire pour faire une recommandation

La classification	Avantages	Inconvénients
Recommandation à base du contenu	• La qualité croit avec le temps • Pas besoin d'information sur les autres usagers • Prendre en considération les gouts uniques des usagers • Possibilité de recommander de nouveaux items ou même des items qui ne sont pas populaires	• Nouvel usager • Nouvel item • L'analyse du contenu est nécessaire pour faire une recommandation
Recommandation collaborative	• Ne demande aucune connaissance sur le contenu de l'item ni sa sémantique • La qualité de la recommandation peut être évaluée • Plus les nombre d'usagers est grand plus la recommandation est meilleure	• Utilisateur avec un goût unique peut pas obtenu une bonne recommandation • Nouvel user • Démarrage à froid • Nouvel Item • Nouvel usager • Problème de confidentialité • La complexité : dans les systèmes avec un grand nombre d'items et d'utilisateurs, le calcul croît linéairement

TABLE 1.2 – Les avantages et les inconvénients des méthodes de recommandation

1.7 Architecture du système de recommandation

Un système de recommandation cherche à prédire la préférence qu'un usager attribuerait à un objet (livre, musique, film...) ou à un élément social (personne, groupe, communauté) qu'il n'avait pas encore considéré.

Pour cela il requiert généralement trois étapes :

1. La première consiste à recueillir de l'information sur l'usager.
2. La deuxième consiste à bâtir une matrice ou un modèle utilisateur contenant l'information recueillie.
3. La troisième consiste à extraire à partir de cette matrice une liste de recommandations.

1.7.1 La collecte d'information

Pour être pertinent, un système de recommandation doit pouvoir faire des prédictions sur les intérêts des usagers. Il faut donc pouvoir collecter un certain nombre de données sur ceux-ci afin d'être capable de construire un profil pour chaque usager. Cette phase peut être faite sous deux façons, soit explicitement ou implicitement La Collecte de données explicite (Filtrage actif) : repose sur le fait que l'usager indique explicitement au système ses intérêts en demandant par exemple à l'usager de commenter, aimer ou encore ajouter comme favoris des contenus (objets, articles...) qui l'intéressent. On utilise souvent une échelle de votes (ratings)allant de 1 étoile (je n'aime pas du tout) à 5 étoiles (j'aime beaucoup) qui sont ensuite transformées en valeurs numériques afin de pouvoir être utilisées par les algorithmes de recommandation.

La Collecte de données implicite (Filtrage passif), repose sur une observation et une analyse des comportements de l'usager effectué de façon implicite dans l'application qui embarque le système de recommandation. Le tout se fait en "arrière-plan" (en gros sans lui rien demander). Par exemple, il obtient la liste des éléments que l'usager a écoutés, regardés ou achetés en ligne. Il analyse la fréquence de consultation d'un contenu par l'usager, le temps passé sur une page et an son réseau social.

1.7.2 Modèle Utilisateur

Pour qu'un système de recommandation donne des prédictions sur les intérêts d'un utilisateur, il doit apprendre un modèle de l'utilisateur. Un modèle de l'uti-

	Avantages	Inconvénients
Collecte Explicite	• Capacité à reconstruire l'historique d'un individu • Capacité à éviter d'agréger une information qui ne correspond pas à cet unique utilisateur (plusieurs personnes sur un même poste).	les informations recueillies peuvent contenir **un biais dit de déclaration**.
Collecte Implicite	• Aucune information n'est demandée aux utilisateurs, toutes les informations sont collectées automatiquement. • Les données récupérées sont a priori justes et ne contiennent pas de biais de déclaration.	• Les données récupérées sont plus difficilement attribuables à un utilisateur et peuvent donc contenir des **biais d'attribution** (utilisation commune d'un même compte par plusieurs utilisateurs). • Un utilisateur peut ne pas aimer certains livres qu'il a acheté, ou il peut l'avoir acheté pour quelqu'un d'autre.

TABLE 1.3 – Avantages et Inconvénients de la collecte explicite et implicite

lisateur se présente généralement sous forme de matrice. On peut le représenter comme un tableau qui contient des données recueillies sur l'utilisateur associées aux produits disponibles sur le site web. Il doit être représenté d'une manière telle que les données peuvent être jumelés aux éléments de la collection. Les éléments que les utilisateurs ont vu dans le passé sont importantes, mais d'autres informations telles que le contenu des articles, la perception des utilisateurs des articles ou des informations sur les utilisateurs eux-mêmes peuvent être utilisés pour construire un profil usager. Ces données doivent être présentées d'une manière telle qu'ils peuvent être utilisés pour différencier les documents sur des sujets différents. Les intérêts des utilisateurs en général ne restent pas les mêmes, mais changent avec le temps. Les données dans le modèle de l'utilisateur doivent donc être ajustées en permanence de sorte qu'il demeure en conformité avec les intérêts de l'usager

1.7.3 Liste de recommandations

Pour extraire une liste de suggestions à partir d'un modèle utilisateur, les algorithmes utilisent la notion de mesure de similarité entre objets ou personnes décrits par le modèle utilisateur. La similarité a pour but de donner une valeur ou un nombre (au sens mathématique du terme) à la ressemblance entre deux choses. Plus la ressemblance est forte, plus la valeur de la similarité sera grande. A l'inverse, plus la ressemblance est faible, et plus la valeur de la similarité sera petite. On verra plus tard dans le dossier quelques exemples.

1.8 évaluation des systèmes de recommandation

Pour évaluer un système de recommandation deux approches sont possibles. Une évaluation online et une évaluation offline. Dans une évaluation online, le système de recommandation est testé par des utilisateurs réels à l'aide d'une application réelle. Ce type d'évaluation a plusieurs avantages. Il permet au système de donner des résultats très fiables, de mesurer la performance de l'application réelle dans un contexte d'utilisation réelle, mais il induit également des problèmes et des difficultés. Tout d'abord, un grand nombre d'utilisateurs est nécessaire pour obtenir des résultats pertinents. Deuxièmement, puisque les utilisateurs réels sont impliqués dans l'expérience, les essais ont un impact sur leur expérience avec le service, ce qui pourrait être indésirable. Troisièmement, les évaluations online ne sont pas applicables à des applications ou des services avant leur lancement.

Puisque les évaluations online avec de vrais utilisateurs sont pour la plupart coûteux et risqué, l'évaluation offline est souvent utilisée comme une méthodologie pour des expérimentations rapides et bon marché [Jannach et al, 2010]. Ce type d'évaluation utilise des ensembles de données de comportement de l'utilisateur historique pour évaluer les nouveaux algorithmes de recommandation ou les paramètres de l'algorithme. L'ensemble de données est divisé en deux ensembles disjoints. L'ensemble d'apprentissage et l'ensemble de test. L'ensemble d'apprentissage représente les interactions que les utilisateurs ont déjà effectuées, et est utilisée comme entrée pour le recommandeur. L'ensemble de test est inconnus pour le système de recommandation et doit être prédit. Le gros avantage des évaluations offline est que de nombreux tests peuvent être effectués dans un court temps, à moindre coût, et sans la nécessité de véritables utilisateurs. Cependant, l'évaluation offline présente aussi des inconvénients. Par exemple, il est difficile de garantir que la méthodologie d'évaluation capte vraiment le comportement de l'utilisateur réel. Dans une évaluation online, les recommandations influent sur le comportement de l'utilisateur, mais l'évaluation offline prévoit simplement un comportement de

l'utilisateur sans l'influence de la recommandation. En conséquence, il est difficile d'estimer le véritable changement dans le comportement de l'utilisateur provoqué par les recommandations.

Pour évaluer la qualité d'un système de recommandation, différents attributs qualitatifs peuvent être mesurées. Le tableau suivant (tableau 1.5) définit quelques un.

La qualité	La description
La précision (Accuracy)	La précision reflète la mesure dans laquelle les recommandations correspondent aux véritables préférences de l'utilisateur. la précision des mesures attribut comment le Système de recommandation peut prédire les éléments que l'utilisateur sélectionnera, comment l'utilisateur va évaluer ces éléments, ou comment l'utilisateur va classer ces articles compte tenu de son / ses préférences personnelles
La diversité (Diversity)	La méthode la plus explorée pour mesurer la diversité dans la liste de recommandation est item-item similarity, comme on peut la mesurer par le calcul de la somme, moyenne, minimum, ou la distance maximale entre les paires d'éléments. Sinon, nous pourrions mesurer la diversité pour chaque élément qui est ajouté à la liste de recommandation de la diversité du nouvel élément dans les articles déjà dans la liste.
La nouveauté (Novelty)	Une nouvelle recommandation est une suggestion pour un élément que l'utilisateur n'était pas au courant de, et que est découverte en explorant les recommandations. Par exemple, si un utilisateur ne connaît pas la sortie d'un nouveau film avec son acteur préféré, alors cet article est une recommandation nouvelle.
La confiance (Trust)	La confiance de l'utilisateur dans le système de recommandation est une mesure importante pour un service efficace. Elle se réfère à la mesure dans laquelle les utilisateurs croire qu'ils aimeront vraiment des recommandations.
L'utilité (Utility)	La plupart des sites de e-commerce évaluent un système de recommandation en termes de (augmentation) des ventes ou le bénéfice qu'ils font... Dans les études de l'utilisateur, l'utilitaire peut être évaluée par un questionnaire ou une interview.

TABLE 1.4 – Les attributs qualitatifs pour évaluer la qualité d'un système de recommandation

La qualité	La description
La satisfaction de l'usager (user satisfaction)	C'est l'objectif final dans l'optimisation d'un système de recommandation, elle se réfère à l'expérience globale de l'usager dans un système de recommandation. La satisfaction de l'usager doit être mesurée par une interview ou un questionnaire
Le risque (Risk)	Dans certains cas, les recommandations peuvent être associées à un risque potentiel, par exemple, un site e-commerce qui permet les achats doivent être retournés à la charge du site. Dans ce scénario, recommandant un mauvais item encourt un coût de livraison (aller et retour).
Le secret (Privacy)	Le système de recommandation de crée un profil personnel, suit le comportement de l'utilisateur, et déduit les préférences personnelles. Ce sont des données sensibles sur l'usager, qui doivent être stockées et traitées en toute sécurité afin de protéger la vie privée des usagers.
La robustesse (Robustness)	La robustesse est une caractéristique du système de recommandation qui indique le degré de stabilité des recommandations dans le cas où de fausses informations sont passées.
Adaptabilité (Adaptivity)	L'Adaptabilité est la caractéristique des systèmes de recommandation indiquant à quelle vitesse les recommandations s'adaptent aux changements dans les articles ou les tendances.

TABLE 1.5 – Les attributs qualitatifs pour évaluer la qualité d'un système de recommandation

1.9 Conclusion

L'étude réalisée dans ce chapitre nous a permis d'avoir une idée claire sur les systèmes de recommandations. Les systèmes de recommandation sont considérés comme étant un sous-ensemble des systèmes hypermédias adaptatifs proposant une solution au problème de surcharge d'information par proposition de recommandations d'items. Le problème de recommandation d'items à un utilisateur a été formalisé et les différents types de systèmes de recommandation et leurs méthodes ont été exposés. Tous les systèmes de recommandation existants emploient une ou plusieurs techniques de base : collaborative, démographique, basée sur la connaissance, sur le contenu, sur l'utilité. Une enquête de ces techniques montre qu'ils ont des avantages complémentaires et des inconvénients. En conséquence, une incitation à la recherche dans les systèmes de recommandation hybrides qui combinent des techniques pour améliorer les performances est recommandée.

Chapitre 2

Web sémantique et Linked data

2.1 Introduction

Au début des années 1990, on a commencé à émerger une nouvelle façon d'utiliser l'Internet pour lier un ensemble des documents. Il a été nommé le World Wide Web. Le Web a commencé comme une collection de documents publiés en ligne accessibles à un endroit du Web identifié par une URL telle que vous pourriez naviguer d'un document à un autre. Ces documents contiennent souvent des données sur les ressources du monde réel qui sont principalement écrites en langage humain et ne peuvent pas être comprises par les machines.

La quantité de connaissances structurées disponibles sur le web ne cesse de croître pour attendre le développement actuel. Cet aspect de la vision originale de Tim Berners-Lee s'est accéléré beaucoup au cours des dernières années et a vu l'émergence du Web de données liées. Le Web de données liées n'est pas fondé seulement sur la connexion des ensembles de données, mais aussi sur la liaison de l'information au niveau d'une seule déclaration.

Le Web des données vise à permettre l'accès à ces données, en les rendant disponibles dans des formats lisibles par machine et en les connectant à l'aide d'identificateurs de ressources uniformes (URI), permettant ainsi aux humains et aux machines de collecter ces données et de les assembler afin de les utiliser dans tous les domaines (pour autant que la licence le permette). Dans la terminologie de web sémantique, le terme de données liées est utilisé pour décrire une méthode d'exposition et la connexion des données sur le Web à partir de différentes sources. Actuellement, le Web utilise les liens qui permettent aux utilisateurs de passer d'un document à un autre. L'idée derrière les données liées est que les humains ou les machines trouvaient des données liées sur le Web qui n'était pas lié précédemment.

2.2 Historique

Le Web est né au CERN, le centre européen de recherche nucléaire, à la fin des années 1980 porté, entre autres, par Tim Berners-Lee chercheur dans ce laboratoire. Il part d'un constat simple : l'absence de cadre d'interopérabilité pour échanger dans un espace de machines en réseau les documents et les données contenus dans les ordinateurs des chercheurs du CERN. Pour régler ce problème, Tim Berners-Lee propose la mise en place d'un dispositif technologique pour mettre à disposition, lier et partager des documents sur un réseau de machines connectées composé de quatre briques technologiques :

- Un protocole de communication, HTTP, basé sur le protocole TCP/IP, c'est-à-dire Internet
- Un mécanisme d'identification, URL, qui permet d'atteindre un document sur un réseau distribué de machines
- Un principe de mise en relation des documents, l'hypertexte créé à l'issue de la seconde guerre mondiale par Vanevar Bush et adapté à l'informatique par Ted Nelson au milieu des années 1960
- Un langage d'encodage des documents, HTML basé sur SGML, une norme de structuration hiérarchique de l'information

Si ces quatre briques technologiques sont à l'origine du Web de documents que nous connaissons aujourd'hui, la proposition initiale de Tim Berners-Lee contenait également la mise en relation des données structurées contenues dans les bases de données des chercheurs. Néanmoins, de ce point de vue, les propositions étaient alors moins concrètes.

Un an après la mise à disposition du premier navigateur Web graphique, Mosaïc, a lieu en septembre 1994 la première conférence WWW au CERN à Genève au cours de laquelle la création du W3C est annoncée. À cette occasion, Tim Berners-Lee dresse les futures directions du W3C et démontre le « besoin de sémantique pour le Web ». Il montre alors en quoi la vision habituelle de l'hypertexte, à savoir la mise en relation de documents par des liens, doit être dépassée pour permettre aux machines de relier automatiquement les données sur le Web aux choses du monde réel. Ambitieuse, l'idée n'en rencontre pas moins des problématiques existantes, en particulier dans le domaine de l'intelligence artificielle.

Lors de suite de cette conférence, outre la mise en place des recommandations nécessaires à la structuration des documents, le W3C lance les premières réflexions dans ce sens. Elles aboutissent à la publication d'un premier brouillon de recommandations en octobre 1997 puis d'un second en avril 1998. La même année, Tim Berners-Lee initie une feuille de route pour le Web sémantique, qui constitue un plan de travail précis des différentes technologies à mettre au point pour le déployer. Dans ce document, il présente le Web sémantique comme une extension du Web de documents qui constituerait une base de données globale à l'échelle du réseau pour permettre aux machines de mieux appréhender les données et aux personnes de coopérer. Cette feuille de route se matérialise par une représentation graphique, le « layer cake », qui montre l'agencement des différentes briques technologiques. Cette représentation est toujours utilisée aujourd'hui, même si les briques ont évidemment évolué.

Par ailleurs, en 1999, il publie le livre Weaving the Web dans lequel il dresse un portrait du Web et les pistes pour son avenir. Les idées du Web sémantique n'en sont évidemment pas absentes.

Les différents travaux engagés depuis 1994 sont présentés pour la première fois au grand public à l'occasion d'un article publié dans la revue Scientific American en mai 2001. écrit par Tim Berners-Lee, Ora Lassila et James Hendler, cet article présente un cas d'utilisation et les différentes technologies nécessaires à son accomplissement. Si cet article permet une introduction pédagogique aux objectifs poursuivis par le Web sémantique, il n'en reste pas moins exploratoire, trop peut-être. De plus, comme James Hendler l'avouera plus tard, il présente le défaut de reprendre certains concepts ou technologies, en particulier le principe des ontologies, qui renvoient aux problématiques de l'intelligence artificielle dont les fantasmes se sont transformés pour le grand public en espoir déçu. Enfin, le mot « sémantique » de par sa polysémie n'aide pas à une compréhension immédiate du concept et des objectifs visés

Malgré les avis dubitatifs et les critiques grandissantes, le W3C continue le travail de normalisation avec la publication de recommandations essentielles : RDFS, OWL et une révision de RDF en 2004, GRDDL en 2007, SPARQL en 2008 et RDFa en 2008 sur lesquelles nous reviendrons plus en détail.

À partir de 2006, deux facteurs vont faire prendre au Web sémantique la direction qui est encore la sienne. Tout d'abord, le Web 2.0 marque l'apparition d'une réflexion dans la mise à disposition des données sur le Web via les Web services,

des principes d'indexation collaborative (folksonomie), mais aussi de la structuration des données d'une page HTML avec le concept des microformats. Soient autant de sujets qui ont trait aux problématiques d'exposition, de structuration et de traitement des données structurées au cœur, également, de la réflexion sur les technologies du Web sémantique.

Par ailleurs, conscient des malentendus engendrés par l'utilisation du mot « sémantique » et des concepts de l'intelligence artificielle, Tim Berners-Lee décrit dans une autre note l'idée du « Linked Data ». Il y rappelle, pour commencer, que le Web sémantique n'a pas vocation uniquement à poser des données dans le Web, mais à relier les données directement entre elles pour qu'une machine ou un humain puisse explorer le Web de données. Il établit quatre règles basées sur les technologies du Web sémantique pour publier sur le Web des données structurées dans un cadre d'interopérabilité commun.

À la suite de cet article et à la vue de l'évolution du Web, deux chercheurs impliqués dans le Web sémantique et issus des domaines de l'intelligence artificielle et de la logique de description vont publiquement reconnaître leur erreur d'appréciation dans leur volonté d'introduire certaines notions complexes dans le Web sémantique. James Hendler parle de « côté obscur du Web sémantique » et avoue que l'introduction d'une logique de description complexe était une erreur stratégique. Chris Welty dans une keynote à ISWC 2007 intitulé « How I was right even when I was wrong » rappelle que l'important dans le Web sémantique, ce n'est pas la sémantique, mais le Web. Cette remise en question issue du bilan des recherches depuis 2001 aboutit à l'aveu de Tim Berners-Lee dans le magazine La Recherche en novembre 2007 :

« Le terme sémantique prête un peu à confusion car la sémantique s'intéresse au sens du langage pour en déduire des constructions logiques. Du coup, certains ont pensé qu'il s'agissait d'un Web qui permettrait par exemple d'effectuer des recherches sur Internet en posant des questions sous forme de phrases, en langage naturel. Or ce n'est pas son but. En fait, nous aurions dû l'appeler dès le départ "Web de données". »

Soutenue par le projet « Linking Open Data » piloté par le W3C, l'idée du Linked Data (qu'on traduira par Web de données) connaît sa plus importante réalisation dès février 2007 avec la création de Dbpedia par deux universités allemandes. Ce projet met à disposition selon les règles édictées par Tim Berners-Lee et, par conséquent, avec les technologies du Web sémantique, les données structurées extraits automatiquement de Wikipedia. En rencontrant une des réussites les

plus médiatiques du Web 2.0, le Web de données acquiert immédiatement une base de travail solide mais aussi une bonne visibilité auprès des spécialistes du Web et de son évolution. Néanmoins, il faudra attendre 2009 et la communication de Tim Berners-Lee à la conférence TED au cours de laquelle il lance son appel « Raw Data Now » pour voir le Web de données atteindre une très large audience. Elle se manifestera en 2010 par l'élaboration du projet de mise à disposition des données gouvernementales britanniques, data.gov.uk, dirigé par Tim Berners-Lee et Nigel Shadboltet basé pour une large partie sur les technologies du Web sémantique.

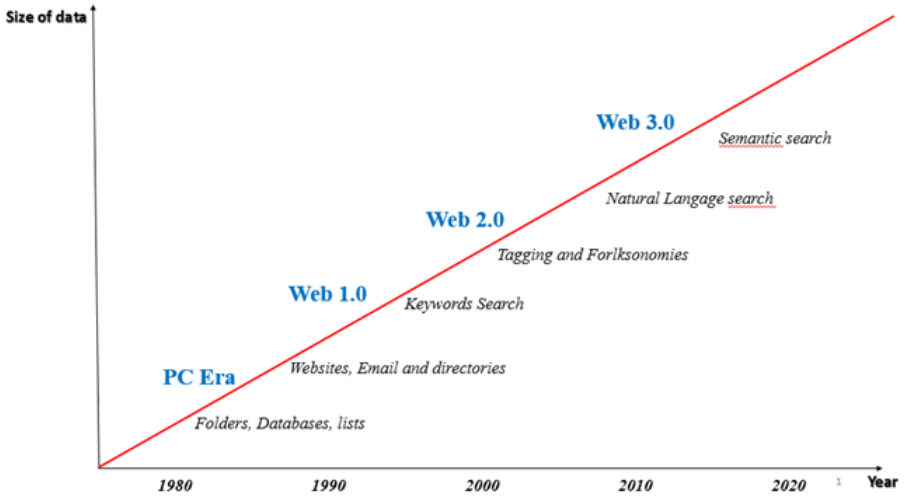

FIGURE 2.1 – Evolution du web

2.3 Les Données ouvertes (open data)

Les données ouvertes représentent un paradigme émergeant du partage des informations, à partir des agences gouvernementales, mais aussi des autres organisations. Elles ont pour but de rendre l'exploitation des données possibles, que ce soit pour un profit social, académique ou économique. Elles sont désignées pour construire une nouvelle communauté de pratique autour du partage d'information et de la réutilisation de celle-ci. Il s'agit de mettre à la disposition de tout le monde et sans restriction d'usage ou de copyright les données brutes produites par la collectivité.

L'Open Data est une démarche qui vise à rendre des données numériques accessibles et utilisables par tous. Pour les collectivités et les organismes publics, elle

consiste à publier sur une plate-forme ouverte des informations : par exemple, la mairie publie des données sur Internet : horaires de bus, géolocalisation des arrêts, hauteur des trottoirs, ce qui permet aux développeurs de réutiliser librement ses données pour créer des applications accessibles à partir d'un téléphone mobile pour faciliter l'accès aux transports aux personnes à mobilité réduite. La mise à disposition des données publiques est une obligation légale. Un cadre juridique strict définit les informations qui peuvent être rendues publiques et celles qui ne le peuvent pas. Les données sensibles et à caractère personnel sont exclues.

En informatique, l'open data est une information structurée publique ou privée et généralement non utilisable par un humain mais interprétable par une machine. Ces informations sont donc exploitées dans les domaines des services web, ou des systèmes d'information interopérables qui réutilisent des fichiers entre eux. De manière générale, la donnée publique / ouverte est une donnée brute, mise à disposition, de tous, sans restriction d'usage. Elle est produite par la collectivité (c'est-à-dire, un internaute, une communauté, etc.). Cette donnée peut être du domaine de l'art, musique, imagerie, photographie, sciences, vidéos, textes.

2.3.1 Pourquoi les Open Government Data ?

Il y a trois raisons principales pour que les données ouvertes doivent être ouvertes :

- **Transparence :**dans un bon fonctionnement, les citoyens de la société démocratique ont besoin de savoir ce que fait leur gouvernement. Pour ce faire, ils doivent pouvoir librement accéder aux données et l'information du gouvernement et de partager cette information avec les autres citoyens. La transparence n'est pas seulement sur l'accès, c'est aussi sur le partage et la réutilisation, pour comprendre la matière, elle doit être analysée et visualisée et cela nécessite que le matériel soit ouvert afin qu'il puisse être librement utilisé et réutilisé.

- **Libérer la valeur sociale et commerciale :** dans l'ère numérique, les données sont une ressource principale pour des activités sociales et commerciales. En ouvrant Les données liées ouverts (Linked open data) les données du gouvernement peuvent aider à conduire la création d'entreprises et des services qui offrent une valeur sociale et commerciale innovante.

- **Gouvernance participative :** une grande partie du temps des citoyens peuvent seulement engager dans leur propre gouvernement sporadiquement peut-être juste à une élection tous les 4 ou 5 ans. En ouvrant les données, les

citoyens sont activés à être beaucoup plus directement informés et impliqués dans la prise de décision. C'est plus que la transparence : il s'agit de faire une société pleine « lecture/écriture », pas seulement de savoir ce qui se passe dans le processus de gouvernance, mais être capable de contribuer.

FIGURE 2.2 – Diagramme de l'open government

2.3.2 Les Caractéristiques des données publiques ouvertes (Open Government Data)

D'après Wikipédia, une donnée ouverte est « une donnée numérique, d'origine publique ou privée, publiée de manière structurée selon une méthodologie qui garantit son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière ». On pourrait résumer en disant que l'Open Data se réfère au monde des données déjà constituées et accessibles et qui deviennent disponibles sous un certain régime juridique et sous un certain format technique qui en permettent ou en facilitent la réutilisation. Une définition plus précise nous est fournie par l'Open Definition de l'Open Knowledge Foundation (http ://opendefinition.org), qui considère qu'une donnée n'est ouverte que s'il est possible de l'utiliser, de la réutiliser et de la redistribuer librement avec, comme seules conditions admissibles, d'une part, l'obligation d'en mentionner la source et, d'autre part, la nécessité de permettre la réutilisation de toute base de donnée dérivée sous les mêmes conditions que la base de donnée originale (share-alike).

Une caractéristique centrale réunit donc le mouvement de l'Open Data : une donnée ouverte implique sa libre réutilisation. Or, cette définition ne rend qu'insuffisamment compte du mouvement et des politiques dont on parle dans le secteur public, ou dans le secteur subventionné par des fonds publics. Surtout on s'aperçoit

que, aussi bien dans la loi que dans la pratique, une donnée publique dite ouverte ne correspond pas toujours à l'intégralité des caractéristiques relevées dans cette définition.

Tout d'abord, dans de nombreux pays y compris en France, la loi garantit un droit aux administrations publiques de maintenir la paternité de leurs données, d'en préserver l'intégrité, et, dans certains cas, de soumettre la réutilisation de ces données au paiement d'une redevance. Ensuite, à l'exception des administrations publiques centrales de l'état, de nombreuses institutions publiques ne sont pas soumises à l'obligation de mettre à disposition leurs données de manière libre et gratuite et sont donc libres d'imposer un certain nombre de conditions ou de restrictions sur les modalités de réutilisation (c'est le cas, par exemple, en France ainsi que dans de nombreux autres pays d'Europe, des collectivités territoriales, des SPIC, des EPIC, des institutions culturelles et de recherche).

Ainsi, une donnée du secteur public au sens large – sauf si elle est dans le domaine public est rarement totalement ouverte (open) et intégralement gratuite (free). Dans le secteur public, l'Open Data pourrait donc être défini comme la mise à disposition libre et (quasi) gratuite de données qui implique la possibilité de les réutiliser de façon la moins contrainte possible. Quasi-gratuite signifie par exemple qu'un certain coût de restitution peut être facturé par ex. pour couvrir les coûts de gestion ou de mise à disposition des données bien que cette redevance ne corresponde en aucune façon au coût réel de production (comme le Journal officiel version papier est quasi-gratuit).

L'idée sous-jacente du principe de gratuité est que, puisque la production des données issues des administrations publiques a été financée, entièrement ou en partie, par l'argent des contributeurs ou contribuables (i.e. les taxes), il serait injuste de demander aux citoyens de contribuer à nouveau à couvrir les coûts de production. Il est cependant acceptable, dans certains cas, de soumettre la réutilisation de ces données à une redevance, qui ne doit toutefois pas dépasser le strict minimum afin de rembourser les coûts supplémentaires qui ont été pris en charge par l'administration afin de rendre ces données accessibles au public.

Ainsi, on peut dire que l'Open Data dans le secteur public est, en fin de compte, une question de degré. Les données sont neutres, mais les politiques gouvernementales ne le sont pas. L'Open Data n'est donc pas un domaine entièrement technique, c'est un domaine ancré dans des cultures différentes. Différents pays ou régions, ainsi que différentes institutions sont susceptibles d'adopter des démarches Open

Data différentes, avec différents degrés d'ouverture selon leurs spécificités et besoins respectifs. On s'intéressera ici à la mise en pratique de ces expériences.

2.3.3 Les données cibles à l'ouverture

La donnée apte à être ouverte dépend directement de son domaine. D'abord, elle doit respecter les caractéristiques citées préalablement. Mais encore une donnée scientifique peut être publiée si elle est sûre, celle-ci peut être ré-exploitée par la suite, par d'autres entités scientifiques qui peuvent la développer ou simplement la réutiliser pour d'autres objectifs.

Toutes les données liées aux collectivités locales doivent être mises à disposition, à l'exception des données dont la publication peut nuire à la sécurité, ou encore celles liées au secret industriel ou commercial. De plus, certaines données se situent à la frontière du public et du privé. C'est d'ailleurs un aspect qu'il faut surveiller. Enfin, certaines données peuvent, en raison de leur caractère sensible, présenter des dangers potentiels en matière de sécurité. D'autre part, un organisme public ne peut ouvrir des données que s'il en est propriétaire. Il devra donc s'assurer que sa démarche est bien conforme au droit de la propriété intellectuelle des agents publics, au droit de la concurrence, aux clauses des marches publics qu'il a conclues

2.3.4 Les cinq étoiles de l'open data

On peut se poser la question de savoir quelle est la déférence entre Linked data et open data, Pour savoir la différence on peut se référer à modèle proposé par Tim Berners-lee dans l'une de ses conférences, où il a attribué un système d'étoiles pour expliquer les différents niveaux d'ouvertures des données.

- Si on met des données sur le web avec une glissance qui autorise à les utiliser, déjà vous faites de l'open data et là c'est une étoile
- Si en plus de cette licence de réutilisation, ce sont des données dans son format qui sont exploitable par machine, par exemple le format Excel et bien on a deux étoiles.
- Si en plus au lieu de choisir Excel, on choisit un format qui n'est pas propriétaire CSV, TXT ou XML, JSON, ODF,... et bien on 3 étoiles
- Si on utilise les standard du web sémantique par exemple le RDF où on utilise des URI pour désigner les données on passe au 4 étoiles

- Si on crée des liens entre les données pour fournir le contexte, on atteint les 5 étoile et c'est ce qu'on l'appelle Linked data

2.4 Le web Sémantique

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"[Berners-Lee , 2001]

Le Web sémantique (techniquement appelé « le Web de données ») permet aux machines de comprendre la sémantique, la signification de l'information sur le Web. Il étend le réseau des hyperliens entre des pages Web classiques par un réseau de lien entre données structurées permettant ainsi aux agents automatisés d'accéder plus intelligemment aux différentes sources de données contenues sur le Web et, de cette manière, d'effectuer des tâches (recherche, apprentissage, etc.) plus précises pour les utilisateurs. Le terme a été inventé par Tim Berners-Lee, Co-inventeur du Web et directeur du W3C, qui supervise l'élaboration des propositions de standards du Web sémantique.

La plupart du temps, lorsque l'on prononce le terme de Web sémantique, on parle des différentes technologies qui se cachent derrière. Parmi les plus connues, on peut citer RDF (Ressource Description Framework) qui correspond à un modèle d'information et les formats d'échanges de données en RDF pour communiquer entre différentes applications (RDF/XML, RDF/JSON, N3, Turtle, N-Triples et d'autres). Dans le domaine du Web sémantique, la sémantique des données est décrite par des ontologies avec des langages prévus pour fournir une description formelle de concepts, termes ou relations d'un domaine quelconque. Ces langages sont RDFS (Ressource Description Framework Schéma) et OWL (Web Ontology Language) Il existe aussi des langages de description des données structurées dans du XHTML afin que des outils effectuent un traitement automatique de ses différentes données. Ces langages sont RDFa et Microformats et, nouvellement arrivés avec HTML 5, Micro data. Ensuite, pour finir avec la liste des technologies, il existe un langage de requête, au même titre que SQL pour les bases de données relationnelles, SPARQL, qui effectue des requêtes, mais sur des triplets RDF. Il en existe d'autres (RQL et RDQL), mais ils sont bien moins utilisés [PLU , 2011]

FIGURE 2.3 – Web sémantique

2.4.1 Les couches du web sémantique

L'ensemble des technologies du Web sémantique est organisé dans une architecture en couches. C'est ce qu'on appelle la « Semantic Web Stack » ou « Pile du Web Sémantique ». Elle constitue la vision du W3C de l'architecture du Web Sémantique. Ainsi, l'ensemble des technologies (langages et protocoles) qu'elle comporte, sont standardisées par le W3C, et font toujours l'objet de recherches et de travaux d'amélioration et de normalisation au sein du W3C (Figure 2.4). Les travaux vi-

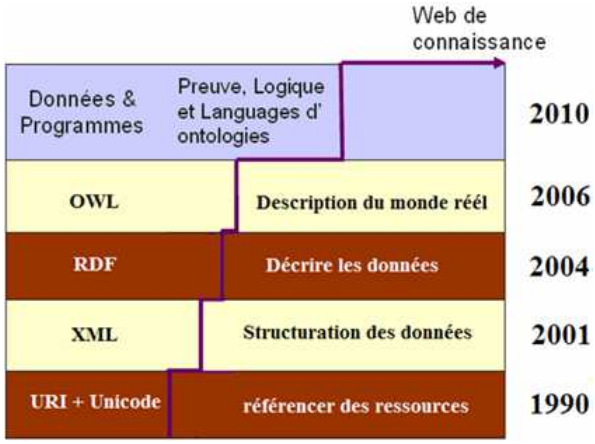

FIGURE 2.4 – Architecture en couches du web sémantique [PLU, 2011].

sant la réalisation du Web sémantique se situent à des niveaux de complexité très différents. Les plus simples utilisent des jeux plus ou moins réduits de métadonnées dans un contexte de recherche d'informations ou pour adapter la présentation des informations aux utilisateurs. Dans ce cas, des langages de représentation simples sont suffisants. Dans les travaux plus complexes mettant en œuvre des architectures sophistiquées, pour permettre par exemple l'exploitation de ressources hétérogènes, des langages plus expressifs et plus formels issus des travaux en représentation et en ingénierie des connaissances, sont nécessaires [Laublet , 2003]

Certaines couches sont déjà implémentées. En effet, les langages et protocoles qui permettent de remplir leurs fonctions existaient déjà ou ont été conçus et créés pour répondre aux spécifications de chaque couche. Le Web sémantique est encore un grand chantier, les couches supérieures ne sont pas encore très évoluées, mais le rôle qu'elles doivent occuper est globalement défini. Chaque couche utilise la couche du dessous et a une fonction bien déterminée dans l'architecture [Polytec, 2012]

La proposition du W3C s'appuie au départ sur une pyramide de langages dont seules les couches basses sont aujourd'hui relativement stabilisées. La Figure 2.4 montre une des versions de l'organisation en couches proposée par le W3C. Deux types de bénéfices peuvent être attendus de cette organisation [Laublet , 2003]

Elle permet une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs. Par ailleurs, si elle est bien conçue, elle doit permettre de disposer du langage au bon niveau de complexité, celle-ci étant fonction de l'application à réaliser.

Un aspect central de l'infrastructure est sa capacité d'identification et de localisation des diverses ressources. Elle repose sur la notion d'URI (Uniform Resource Identifier) qui permet d'attribuer un identifiant unique à un ensemble de ressources, sur le Web bien sûr, mais aussi dans d'autres domaines (documents, téléphones portables, personnes, etc.). Une autre caractéristique de tous ces langages est d'être systématiquement exprimables et échangeables dans une syntaxe XML. Ceci permet de bénéficier de l'ensemble des technologies développées autour d'XML [Laublet, 2003]

Elle permet une approche graduelle dans les processus de standardisation et d'acceptation par les utilisateurs. Par ailleurs, si elle est bien conçue, elle doit permettre de disposer du langage au bon niveau de complexité, celle-ci étant fonction de l'application à réaliser.

Un aspect central de l'infrastructure est sa capacité d'identification et de localisation des diverses ressources. Elle repose sur la notion d'URI (Uniform Resource Identifier) qui permet d'attribuer un identifiant unique à un ensemble de ressources, sur le Web bien sûr, mais aussi dans d'autres domaines (documents, téléphones portables, personnes, etc.). Une autre caractéristique de tous ces langages est d'être systématiquement exprimables et échangeables dans une syntaxe XML. Ceci permet de bénéficier de l'ensemble des technologies développées autour

d'XML [Laublet, 2003]

2.4.1.1 URI et Unicode

L'URI (Uniform Resource Identifier) [Mestiri, 2007] est justement un protocole simple et extensible pour identifier, d'une manière unique et uniforme, toute ressource sur le web. Il s'agit d'un aspect central de l'infrastructure, c'est pour cette raison que cet élément se trouve à la base de l'architecture en couches proposée. La spécification de la syntaxe et de la sémantique de l'URI dérive des concepts introduits par l'initiative de globalisation de l'information dans le World Wide Web, qui date des années 1990.

Dans le cas du web sémantique, l'URI est une séquence de caractères avec une syntaxe restreinte, qui permet d'identifier toute ressource utilisée dans le cadre d'une application web sémantique. On entend par ressource n'importe quel objet ayant une identité, telle qu'un document électronique, une page HTML, un fichier, une image, une vidéo, etc. La notion d'uniformité, ou universalité 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. Par ailleurs, elle permet d'introduire de nouveaux types d'identificateurs de ressources sans interférer avec ceux existants. Enfin, cette uniformité permet la réutilisation de ces identificateurs dans différents contextes [Mestiri , 2007]

Un URI peut être classé, en effet en 3 catégories [Mestiri , 2007] : URL (Uniform Resource Locator) désigne un sous-ensemble d'URI qui identifie les ressources via une représentation de leur mécanisme d'accès, plutôt que par le nom ou autres attributs de cette dernière, comme il en est le cas pour l'URN (Uniform Resource Name). L'URL et l'URN sont donc des cas particuliers d'URI. 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.

2.4.1.2 Le langage XML

La recommandation du W3C pour un langage de balisage extensible, eXtensible markup language (XML) date de 1998, et a assez peu évolué depuis, jusqu'à sa cinquième édition en 2008. Il s'agit d'un format pour structurer les informations au

sein d'un fichier texte. XML est un langage de balisage comme HTML, et fonctionne donc également sur le principe d'insertion de marques, les balises ou tags, dans le flux même du texte d'un fichier, pour délimiter et étiqueter des portions de ce texte. Mais là où HTML propose un ensemble prédéfini et limité de balises destinées avant tout à guider les traitements effectués par un navigateur Web lors de l'ouverture d'une page, XML permet de définir ses jeux de balises et ne préjuge pas de leur finalité ni de leur interprétation. XML se contente d'imposer une syntaxe (quelques règles relativement simples) destinée à mettre en évidence la structure de l'information inscrite dans un fichier [Menon , 2013].

S'il est possible de définir librement la structure d'un document XML (éléments, attributs et contraintes liées à leur position dans le fichier), on peut aussi spécifier des structures type auxquelles on choisira de se conformer. Les mécanismes le permettant sont les définitions de type de document (DTD) et les schémas XML. XML est aujourd'hui un format d'échange de données et de documents dont l'usage s'est beaucoup répandu, bien au-delà du contexte du Web sémantique.
Il s'agit d'un standard stable, perçu comme pérenne, indépendant de tout constructeur de matériel informatique ou fournisseur de logiciel, ce qui a certainement favorisé son adoption [Menon , 2013].

2.4.1.3 Langage RDF et RDF Schéma

RDF (Ressource description Framework)

Resource Description Framework (RDF) [Menon , 2013], est un standard du W3C publié originellement en 1999, dont une nouvelle version a été mise au point en 2004 ; une révision de RDF est en cours, pour y apporter quelques ajustements et clarifications. Remarquons auparavant que sa portée a évolué : la recommandation de 1999 parle d'une « base pour traiter les métadonnées », et d'un « mécanisme de description des ressources » indépendant des applications et ne préjugeant pas de la sémantique attachée à ces descriptions ; en 2004, on définit RDF comme un « cadre pour représenter les informations sur le Web ». Cette évolution montre la prise de conscience du caractère de très grandes généralités de RDF, qui est conçu pour permettre « à n'importe qui de dire n'importe quoi sur n'importe quoi ». Elle traduit aussi le fait qu'une ressource sur le Web peut bien entendu être un document, un texte, une image, mais peut tout aussi bien être l'avatar dans le monde virtuel, via l'attribution d'un URI, d'une personne, d'un objet, d'un événement, d'un lieu ou d'un concept. Le RDF est un modèle conceptuel, servant à encadrer la description de ressources d'une façon simple et non ambiguë [Benayache, 2005]. Ses applications visent initialement le web sémantique, mais elles peuvent s'étendre

plus largement à l'ingénierie des connaissances. Ce modèle est associé à une syntaxe dont le but est de permettre à une communauté d'utilisateurs de partager les mêmes métadonnées pour des ressources partagées. Le RDF repose sur un ensemble de vocabulaires pour écrire les métadonnées, dont les plus connus sont le DC (Dublin Core) et le FOAF (Friend Of A Friend). Dans le web sémantique, un vocabulaire désigne un ensemble de termes utilisés pour décrire des ressources [Mestiri , 2007].

Le Modèle RDF des données

Le premier but de RDF est donc de permettre d'exprimer des métadonnées, c'est-à-dire de décrire des ressources, de façon simple, standardisée et distribuée (une même ressource n'a pas besoin d'être intégralement décrite à un seul endroit). Pour ce faire, on propose un modèle qui repose sur l'assertion RDF, ou déclaration RDF : on attribue à une ressource une propriété munie d'une valeur. Nous empruntons à Bernard Vatant une formulation concise de ce modèle : « La description RDF d'une ressource est donc un ensemble de triplets (sujet, prédicat, objet) où le sujet est la ressource à décrire, le prédicat une propriété applicable à ce sujet, et l'objet une valeur de cette propriété ». RDF est souvent représenté graphiquement comme dans la Figure 2.5. [Menon , 2013]. À un même sujet peut être associé plusieurs couples

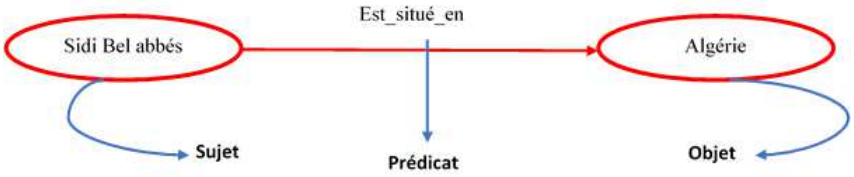

FIGURE 2.5 – Exemple d'une représentation RDF

prédicat/objet. L'objet d'une assertion peut parfaitement devenir le sujet d'une autre assertion. Par ailleurs il peut être également l'objet de plusieurs assertions, de sorte que le modèle permet la construction de graphes, où les sujets et les objets sont des noeuds et où les prédicats sont des arcs qui les relient, orientés du noeud sujet vers le noeud objet [Menon , 2013].

Par exemple, si je souhaite exprimer l'assertion *« une personne qui s'appelle Slimane, habite à 8 rue Amir Abdelkader Sidi Bel Abbes, il a étudié à l'université d'Oran »*, le graphe sera défini comme suite (Figure2.6). Dans cet exemple en remarque les triplets suivants : (personne, habite à, maison) ; (personne, nom, Slimane) ;(Slimane, a étudié, université) ; (université, ville, Oran) ;etc. On remarque que le sujet personne a deux prédicats et deux objets, l'objet Slimane est un sujet

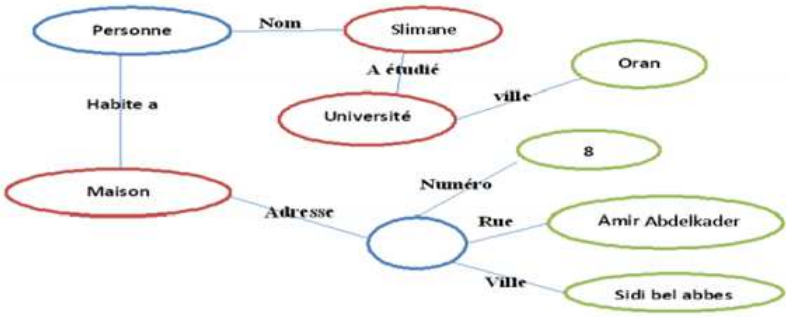

FIGURE 2.6 – Exemple d'un graphe RDF

pour l'objet université et il a comme sujet personne. Dans les assertions RDF (triplets), la ressource (sujet) est identifiée par un URI, de même que la propriété (prédicat). La valeur de la propriété (objet) peut également l'être, mais peut aussi être un littéral (ou valeur atomique), comme une chaîne de caractères ou un nombre.

Comme nous sommes sur le Web, l'utilisation d'URI permet, où que l'on soit, de faire référence aux éléments ainsi identifiés des triplets ; un autre nom proposé par Tim BernersLee en 2007 pour le Web sémantique (Global giant graph) [MENON, 2013]. Par exemple, une partie de la représentation des triplets précédents « une personne qui s'appelle Slimane » forme le graphe suivant :

FIGURE 2.7 – Exemple d'un graphe RDF avec des URI

2.4.1.4 La syntaxe RDF

Comment doit-on exprimer concrètement les métadonnées RDF pour qu'elles puissent être stockées, transmises et traitées en machine ? Autrement dit, quelle

syntaxe ?, quelle notation adopter pour RDF ?, sachant que de nombreuses possibilités existent. Sur ce point, la recommandation RDF ne se prononce pas, mais le W3C a publié en 2004 la recommandation RDF/XML qui, comme son nom l'indique, définit l'expression en syntaxe XML des graphes RDF. Cette approche est apparue initialement comme la plus cohérente, dans la mesure où XML est la base sur laquelle les autres niveaux de l'architecture d'ensemble sont construits . Par exemple, si je souhaite exprimer l'assertion *« une personne qui s'appelle Slimane »* le graphe sera défini comme suite :

```
< ?xml version="1.0" encoding="UTF-8" ?>
<rdf :RDF
      xmlns :foaf="http ://xmlns.com/foaf/0.1/"
      xmlns :rdf="http ://www.w3.org/1999/02/22-rdf-syntx-ns"
      xmlns :rdfs="http ://www.w3.org/2000/01/rdf_schema"
      xmlns :dc="http ://purl.org/dc/terms/">
<rdf :Description rdf :about="http ://www.w3.org/People/slimene/cardi">
     <rdf :type rdf :resource="http ://xmlns.com/foaf/0.1/Person">
      <foaf :name>Slimane<foaf/ :name>
</rdf :Description>
</rdf :RDF>
```

Mais la syntaxe RDF/XML, bien que largement utilisée, a fait l'objet des critiques parfois virulentes chez les développeurs : limitée dans ce qu'elle peut exprimer, trop bavarde, trop complexe, opaque et ne permettant pas de repérer facilement les sujets, prédicats et objets, difficiles à manipuler avec les outils XML existants, etc. C'est pourquoi le W3C met actuellement au point une autre notation, Turtle, déjà utilisée dans un certain nombre de contextes, qui se veut compacte et « naturelle », et qui est actuellement au stade de projet de recommandation (juin 2012). On peut insérer des métadonnées formulées selon RDF/XML ou Turtle à des pages Web (écrites en HTML ou XHTML), mais c'est assez laborieux et le résultat est un peu lourd ; en particulier, on est alors conduit à répéter en RDF des informations déjà présentes dans le texte (X)HTML. On préférera recourir à RDF in attributes, ou RDFa, une recommandation du W3C dont la version 1.1 a été publiée en juin 2012. Son objectif est de rendre l'insertion de triplets RDF dans des pages Web plus aisée et plus concise, grâce à l'utilisation du mécanisme d'attributs présents dans (X)HTML [Menon , 2013]..

Il existe d'autres formats de sérialisation sont en cours d'utilisation, y compris :

- N-Triples, un format très simple, facile à analyser, basé sur la ligne qui n'est pas aussi compact que Turtle.

- N-Quads, un sur-ensemble de N-Triples, de sérialisation plusieurs graphes RDF.
- JSON-LD, une sérialisation à base de JSON
- N3 ou Notation 3, une sérialisation non standard qui est très similaire à Turtle, mais il a quelques fonctionnalités supplémentaires, comme la possibilité de définir des règles d'inférence.

La sérialisation en N3 des triplets exprimés dans l'exemple précédent est la suivante :

```
@prefix rdfs :http ://www.w3.org/2000/01/rdf_schema >
@prefix foaf :http ://xmlns.com/foaf/0.1/>
@prefix rdf :http ://www.w3.org/1999/02/22-rdf-syntx-ns>
@prefix dc :<http ://purl.org/dc/terms/>
<http ://www.w3.org/People/slimene/cardi>
rdf :type foaf :Person;
foaf :name "slimane";
```

2.4.1.5 RDFa

RDFa [Heath and Bizer , 2011] est un format de sérialisation qui intègre les triplets RDF dans les documents HTML. Les données RDF ne sont pas intégrées dans les commentaires dans le document HTML, comme c'était le cas avec quelques premières tentatives de mélanger RDF et HTML, mais il est entrelacé dans le modèle objet de document HTML (DOM). Cela signifie que le contenu existant au sein de la page peut être marqué avec RDFa en modifiant le code HTML, exposant ainsi les données structurées sur le Web. RDFa est populaire dans des contextes où les éditeurs de données sont capables de modifier des modèles HTML, mais ont relativement peu de contrôles supplémentaires sur l'infrastructure de l'édition.

Par exemple, de nombreux systèmes de gestion de contenu permettront aux éditeurs de configurer les modèles HTML utilisés pour exposer différents types d'informations, mais peuvent ne pas être suffisamment flexibles pour supporter 303 redirections et la négociation de contenu HTTP.

Lors de l'utilisation de RDFa pour publier les données liées sur le Web, il est important de maintenir la distinction non ambigu entre les objets du monde réel décrits par les données et le document HTML + RDFa qui incarne ces descriptions. Ceci peut être réalisé en utilisant le RDFa about = attribuent à attribuer les références URI aux objets du monde réel décrits par les données. Si ces URI

sont d'abord définis dans un document RDFa, ils suivront le modèle l'URI hachage [Heath and Bizer , 2011].

2.4.1.6 Schéma et les vocabulaires RDF

RDF Schéma ou RDFS est un langage extensible de représentation des connaissances. Il appartient à la famille des langages du Web sémantique publié par le W3C. RDFS fournit des éléments de base pour la définition d'ontologies ou vocabulaires destinés à structurer des ressources RDF.

Ils permettent de déclarer des propriétés et des classes (ce qui en fait des ressources identifiées par des URI) et de spécifier la signature (en termes de domaine et de portée) des propriétés. RDFS offre également un mécanisme de construction de taxonomies de propriétés et de classes, pour faire de telle propriété une « sous-propriété » de telle autre, ou de telle classe une sous-classe de telle autre. On peut ainsi indiquer, par exemple, que la propriété « auteur » s'applique à la classe de ressources « oeuvre », qui a pour sous-classe « livre » et prend ses valeurs dans la classe « personne ».[Menon , 2013]

RDFS est lui-même un vocabulaire RDF : il définit l'ensemble fermé des propriétés et des classes de ressources utilisables pour le domaine d'application consistant à spécifier les autres vocabulaires. Un vocabulaire spécifié selon RDFS se présente donc formellement comme un graphe RDF [Menon , 2013].

La première version de RDFS a été proposée en mars 1999, et la recommandation finale publiée par le W3C en février 2000. Les composants principaux de RDFS sont intégrés dans un langage d'ontologie plus expressif, OWL.
De nombreux vocabulaires RDF ont été définis dans divers domaines d'application. En voici quelques exemples :

- Le projet Friend of a Friend (FOAF) a défini un vocabulaire RDF permettant de décrire des personnes : leur identité, leurs centres d'intérêt, leurs activités, les relations qu'ils entretiennent, etc.
- Les métadonnées telles que définies par le Dublin core metadata initiative (DCMI) font également l'objet d'une recommandation pour leur expression en RDF, émise par le même organisme.
- Le Data cube vocabulary, vocabulaire pour les cubes de données, vise à permettre la publication en RDF de données statistiques, vues comme des en-

sembles de valeurs distribuées sur un certain nombre de dimensions et accompagnées de métadonnées

- Simple knowledge organization system (SKOS), système simple d'organisation des connaissances, se veut un modèle de données commun à divers types de langages documentaires, tels que thésaurus, classifications, listes de vedettes-matière conçuent pour en faciliter la publication, leur mise en commun et leur exploitation dans le cadre du Web sémantique. En 2009, ce modèle a donné lieu à une recommandation du W3C qui comporte notamment un vocabulaire RDF

Grâce à un mécanisme similaire à celui des espaces de noms en XML, un même jeu de métadonnées RDF peut utiliser des propriétés et des classes définies dans différents vocabulaires. Il est possible d'utiliser le langage de requête SPARQL pour les atteindre à travers le Web.

2.4.1.7 Langage SPARQL

SPARQL (prononcé sparkle en anglais) est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles à travers Internet. Son nom est un acronyme récursif qui signifie SPARQL Protocol and RDF Query Language. SPARQL est l'équivalent de SQL, car comme en SQL, on accède aux données d'une base de données via ce langage de requête alors qu'avec SPARQL, on accède aux données du Web des données. Cela signifie qu'en théorie, on pourrait accéder à toutes les données du Web avec ce standard. L'ambition du W3C est d'offrir une interopérabilité non pas seulement au niveau des services, comme avec les services Web, mais aussi au niveau des données structurées ou non qui sont disponibles à travers l'Internet. Ce standard a été créé par le groupe de travail DAWG (RDF Data Access Working Group) du W3C (Consortium World Wide Web). SPARQL est considéré comme l'une des technologies clés du Web sémantique et le 15 janvier 2008, la version 1.0 est devenue une recommandation officielle du W3C. La version 1.1 permettra d'enregistrer des données et de fusionner des données de sources différentes. La version 1.1 est devenue depuis le 21 mars 2013 une recommandation. Les implémentations de SPARQL au sein de triples stores se multiplient. « SPARQL fera une énorme différence », selon Tim Berners-Lee dès mai 2006. Un exemple de requête sparql où on veut trouver toutes les personnes que connaît Tim Berners-Lee.

Les deux inconnus sont X qui correspond à l'URI des ressources reliés à Tim Berners-Lee parle prédicat « foaf :knows » et Y qui correspond au littéral relié à X par le prédicat « foaf :name ». Le résultat de cette requête sera :

?X	?Y
<http ://www.ivan-herman.net/foaf.rdf#me	"Ivan Herman"
<http ://dbpedia.org/resource/Tim$_B ray$	"Tim Bray"
<http ://danbri.org/foaf.rdf#danbri	"Dan Brickley"
<http ://lassila.org/ora.rdf#me	"Ora Lassili"
<http ://www.cs.umd.edu~hendler/2003/foaf.rdf#jhender>	"James Hendler"

TABLE 2.1 – Extension du Web de Documents vers le Web Sémantique

2.4.2 Technologie de web sémantique

En revenant à la définition du Web Sémantique, nous pouvons comprendre qu'il dépend directement de cette vision de la machine interprétant le contenu d'un texte. Cette même vision est née d'un syllogisme simple : Les documents web ont un sens ; les documents web sont compréhensibles par un ordinateur ; un ordinateur est capable de comprendre le sens d'un document. Naturellement, ce syllogisme n'apparaît simple que si l'on possède des bases en Web Sémantique. Toutefois, la compréhension de ces piliers mène à une véritable assimilation de ce qu'est le Web Sémantique.

- Les Métadonnées
- Les Ontologies
- La logique de raisonnement

2.4.2.1 Les métadonnées

Une métadonnée est « une donnée sur une donnée ». Cette définition est un peu vague voire ambiguë, et elle est comprise de manières différentes par différentes communautés qui conçoivent, créent, décrivent, préservent et utilisent des systèmes d'information et des ressources. Par exemple, dans certains cas, la donnée sur laquelle la métadonnée porte est considérée comme ayant le même statut de données formalisées, traitable par un système informatique, dans d'autres, la donnée n'est qu'interprétable par un être humain, et seule la métadonnée en permet le traitement automatique.

Une métadonnée [Morineau , 2013] peut être externe à la ressource qu'elle décrit (cas d'une microfiche par exemple, d'une notice dans un catalogue) ou interne (balise méta d'une page web, description d'une image web...) notamment dans le cas de données informatiques. Il existe plusieurs types de métadonnées :

- Des métadonnées de gestion, permettant d'accéder au document (auteur, titre, date de création, date de modification, langue)
- Des métadonnées de description, pour en comprendre le contenu (sujet, description)
- Des métadonnées de préservation, pour garantir la pérennité de l'accès et de la compréhension du document (droits, format du fichier, source, résolution, relation, couverture, etc.).

Les métadonnées sont la carte d'identité d'un document. Elles permettent de l'identifier, de le décrire, d'expliquer l'origine de sa création, son utilité et ses destinataires. Au-delà de cette seule description, elles facilitent la recherche et le partage des ressources, la gestion de collections, leur préservation autant que la gestion des droits et l'authentification des documents. L'usage des métadonnées relève d'une pratique ancienne dans les bibliothèques et les centres de documentation, habitués à normaliser le signalement et le contenu des documents. Ainsi, les fiches cartonnées normalisées en 1954 sous la référence ISBD (International standard bibliographic description) ont progressivement fait place à des notices bibliographiques comme le format MARC (Machine-readable cataloging) utilisé pour la description des ouvrages, MARC ISBD(S) (International Standard Bibliogra phic Description for Serials), pour la description des publications en série, puis à des schémas spécifiques de description des ressources numériques, à l'instar du langage Dublin Core.

Les ressources à décrire étant variées, chaque métier s'est ainsi doté de son propre langage de description (EAD (Encoded Archival Description) pour les archives, LOM (Learning Object Metadata) pour les ressources pédagogiques, RKMS (Recordkeeping Metadata Schema) pour les ressources audio ou encore CIMI consortium (Computer Interchange of Museum Information) pour les ressources muséographiques). Soulignons enfin que les métadonnées sont à la base du Web sémantique qui les définit et les utilise dans le cadre du modèle Resource Description Framework (RDF). RDF offre donc une architecture souple qui permet d'ajouter des descriptions sous forme de triplets dans une vision dynamique et évolutive des métadonnées [Morineau , 2013].

Avec l'émergence de pratiques d'indexation collaborative (folksonomies, identification de documents iconographiques ou audiovisuels, correction collaborative d'OCR, etc.), on parle aussi de « métadonnées sociales » pour insister sur l'enrichissement et l'amélioration de la description des collections et donc de l'accès des utilisateurs à ces collections. Un autre aspect de cette appropriation des données transparaît dans l'émergence du mouvement open data qui tend à faire de la

problématique de la réutilisation des informations du secteur public un des enjeux majeurs des politiques culturelles d'aujourd'hui.

2.4.2.2 Les Ontologies

Ontologie est une branche de la métaphysique qui s'intéresse à l'existence, à l'être en tant qu'être et aux catégories fondamentales de l'existant. En effet, ce terme est construit à partir des racines grecques « ontos » qui veut dire ce qui existe, l'être, l'existant et « logos » qui veut dire l'étude, le discours, d'où sa traduction par « l'étude de l'être » et par extension de l'existence [Benhmidi , 2011]

En informatique, L'ontologie [PLU, 2011] est la base de ce que l'on appelle la représentation des connaissances. Ce domaine est né de la volonté des chercheurs de représenter diverses connaissances du monde actuel, de façon à ce qu'elles soient utilisables par des ordinateurs, pour qu'ils puissent effectuer des raisonnements sur ces connaissances. Ces connaissances sont exprimées sous forme de symboles auxquels on donne une « sémantique » (un sens). Imaginons la problématique suivante : vous voulez interroger une base de données contenant diverses ressources (textes, images, vidéos...) et une requête (question ou mot(s) clé(s)),

comment trouver les données qui se trouvent dans cette base qui correspondent à cette requête ?. Par exemple tapez dans votre moteur de recherche préféré les mots suivants : « ordinateur portable » puis « lap top ». Vous pouvez vous apercevoir que les résultats ne sont pas du tout les mêmes, alors que, vu que les deux mots représentent la même chose, on pourrait s'attendre à trouver les mêmes réponses. Que se passe-t-il ? En fait, le moteur de recherche compare des mots sans prendre en compte leur sémantique (sens). Il exécute uniquement une recherche strictement syntaxique et donc sans réflexion car « ordinateur portable » et « laptop » représentent le même concept (la même chose), que nous appellerons maintenant des classes pour respecter la terminologie du Web sémantique. Plus précisément, on peut dire que la relation de spécialisation sur les classes n'est pas gérée. Par exemple, « notebook » est une spécialisation de la classe générale « laptop ». Ainsi, pour raisonner, il ne faut plus se baser sur les mots mais sur les classes. Mais que signifie raisonner ? Raisonner c'est utiliser sa raison pour démontrer quelque chose. C'est un terme très souvent employé en intelligence artificielle.

Type d'ontologie

Il existe de nombreuses sortes d'ontologies, destinées à des utilisations très variées. [Benhmidi , 2011] distingue six types d'ontologie :

1. ***Ontologie de représentation de connaissances :***
   Modélise les représentations primitives utilisées pour la formalisation des connaissances sous un paradigme donné. Par exemple, une ontologie sur le formalisme des Topic Maps comportera les concepts : Topic, Type de Topic, Association, Occurrence, Type occurrence.

2. ***Ontologie de haut niveau / supérieure (Top-level / Upper-model :)***
   Elle exprime des conceptualisations valables dans différents domaines. Elle décrit des concepts très généraux comme l'espace, le temps, la matière, les objets, les évènements, les actions, etc. Ces concepts ne dépendent pas d'un problème ou d'un domaine particulier et doivent être, du moins en théorie, consensuels à de grandes communautés d'utilisateurs. Ce type d'ontologies est fondé sur la théorie de la dépendance. Son sujet est l'étude des catégories des choses qui existent dans le monde. Haute abstraction telle que les entités, les évènements, les états, les processus, les actions, le temps, l'espace, les relations, les propriétés, etc. Des exemples d'ontologies de haut niveau sont Dolce ou Sumo.

3. ***Ontologie Générique (Generic ontology :)***
   Elle est appelée également noyau ontologique, modélise des connaissances moins abstraites que celles véhiculées par l'ontologie de haut niveau, mais assez générales néanmoins pour être réutilisées à travers différents domaines. Cette ontologie inclut un vocabulaire relatif aux choses, évènements, temps, espace, causalité, comportement, fonction, etc.

4. ***Ontologie du domaine (Domain ontology) :***
   Cette ontologie exprime des conceptualisations spécifiques à un domaine, elle est pour plusieurs applications de ce domaine. Elle fournit les concepts et les relations permettant de couvrir les vocabulaires, activités et théories de ces domaines. Selon Mzoguchi, l'ontologie du domaine caractérise la connaissance du domaine ou la tâche est réalisée. Par exemple, dans le contexte du e-Learning, le domaine peut être celui de formation.

5. ***Ontologie de Taches (Task ontology) :***
   L'ontologie de taches fournit un vocabulaire systématisé des termes employés pour résoudre des problèmes liés aux taches qui peuvent être ou non du même domaine. Elle fournit un ensemble de termes au moyen desquelles nous pouvons décrire généralement comment résoudre un type de problème. Elle

inclut des noms, des verbes et des adjectifs génériques dans les descriptions de taches.

6. ***Ontologie d'application (Application ontology) :***
   C'est l'ontologie la plus spécifique, elle contient des concepts dépendants d'un domaine et d'une tache particulier, elle est spécifique et non réutilisable. Ces concepts correspondent souvent aux rôles joués par les entités du domaine lors de l'exécution d'une certaine activité. Il s'agit donc ici de mettre en relation les concepts liés à une tache particulière de manière à en décrire l'exécution.

Ontologie OWL

L'origine du langage OWL est Le World Wide Web Consortium (W3C) qui a mis sur pieds, en Novembre 2001, le groupe de travail « WebOnt », chargé d'étudier la création d'un langage standard de manipulation d'ontologies web. Le premier Working Draft *«OWL Web Ontology Language 1.0 Abstract Syntax »* paraît en Juillet 2002 et, au final, OWL devient une recommandation du W3C le 10 Février 2004 ; La plupart des systèmes qui utilisent actuellement DAML, OIL et DAML+OIL (langages prédécesseurs d'OWL) sont en train de migrer vers OWL.

Le langage d'ontologie Web OWL est conçu pour décrire et représenter un domaine de connaissance spécifique, en définissant des classes de ressources ou d'objets et leurs relations ; ainsi que de définir des individus et affirmer des propriétés les concernant et de raisonner sur ces classes et individus dans la mesure où le permet la sémantique formelle du langage OWL. OWL est un standard basé sur la logique de description, il est construit sur RDF et RDFS et utilise la syntaxe RDF/XML. Le langage OWL permet d'étendre les technologies de base (XML, RDF, RDFS) pour apporter :

- Plus d'interopérabilité (équivalences)
- Plus de raisonnements (logique de description)
- Plus d'évolution (intégration d'ontologies).

Les ontologies OWL se présentent, généralement, sous forme de fichiers texte et de documents OWL. Le langage OWL offre trois sous langages d'expression conçus RC pour des communautés de développeurs et d'utilisateurs spécifiques.

2.4.2.3 La logique de raisonnement

Le Web sémantique repose sur la représentation formelle des connaissances diffusées sur le Web afin de permettre l'utilisation de mécanismes de raisonnements

automatiques pour l'accès à ces connaissances. Les logiques de description se sont concentrées sur les raisonnements par classification (recherche des éléments plus généraux ou plus spécifiques qu'un élément donné), l'idée générale étant d'associer des descriptions complexes à des classes ou entités de façon à pouvoir classifier une instance ou une nouvelle classe au sein d'un ensemble de classes précédemment décrit. Elles ont ainsi développé des constructeurs adaptés à cet objectif de description de classes mais se sont révélées peu adaptées au problème de l'interrogation des données, ce qui a nécessité le développement de nouvelles logiques de description moins expressives avec de nouveaux mécanismes de raisonnement. Différentes typologies des raisonnements sont possibles.

1. ***Raisonnement formalisé et non formalisé***
   Un raisonnement est dit formalisé s'il s'énonce dans une langue formelle, obéissant à des règles de syntaxe strictes et évacuant ainsi l'ambiguïté sémantique. Typiquement, les raisonnements mathématiques sont des raisonnements formalisés. Un raisonnement peut également être exprimé en langue naturelle et respecter parfaitement des règles logiques d'inférences. Il existe ainsi des degrés plus ou moins « élevés » de formalisme.

2. ***Raisonnement a priori et a posteriori :***
   Le raisonnement a priori, dit aussi « analytique », recourt souvent à une formalisation logique pour établir une preuve. Il repose surtout sur des principes et sur une analyse conceptuelle.
   À l'opposé des raisonnements a priori, il existe des raisonnements a posteriori reposant sur des « données empiriques ». Celles-ci peuvent être recueillies par expérimentation ou observation. Un raisonnement empirique peut être tout aussi rigoureux qu'un raisonnement analytique.

2.5 Définition des données liées

Pour rendre le Web sémantique ou Web de données une réalité, il est nécessaire de disposer d'un volume important de données disponibles sur le Web dans un format standard, accessible et gérable. En outre, les relations entre les données doivent également être mises à disposition. Cette collection de données interdépendantes sur le Web peut aussi être nommée comme Linked Data.
La notion de données liées correspond à une vision plus technologique des données en relation avec le Web sémantique. Il s'agit d'une collection d'ensembles de données publiés en utilisant les langages du Web sémantique (RDF(S) et OWL), ensembles de données reliés les uns aux autres et interrogeables au moyen du langage SPARQL. Les deux points importants sont l'utilisation des langages du Web

sémantique et les liaisons entre ensembles de données ce qui permet dans une même requête d'interroger plusieurs ensembles.

Le Web des données (Linked Data, en anglais) est une initiative du W3C (Consortium World Wide Web) visant à favoriser la publication de données structurées sur le Web, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un réseau global d'informations. Il s'appuie sur les standards du Web, tels que HTTP et URI - mais plutôt qu'utiliser ces standards uniquement pour faciliter la navigation par les êtres humains, le Web des données les étend pour partager l'information également entre machines. Cela permet d'interroger automatiquement les données, quels que soient leurs lieux de stockage, et sans avoir à les dupliquer. Tim Berners-Lee, directeur du W3C, a inventé et défini le terme Linked Data et son synonyme Web of Data au sein d'un ouvrage portant sur l'avenir du Web sémantique.

2.5.1 Les principes de données liées

Le terme donnée liée [93] se réfère à un ensemble de bonnes pratiques à mettre en œuvre pour publier et lier des données structurées sur le Web. Ces pratiques ont été introduites par Tim Berners-Lee dans Linked Data, sa note sur l'architecture du Web et sont connues en tant que principes des données liées. Les voici :

- Nommer les éléments avec des URI ;
- Utiliser des URI HTTP, pour que l'on puisse rechercher/consulter ces noms
- Fournir des informations nécessaires sous forme de standards (RDF, SPARQL) lors d'une recherche d'URI
- Inclure des liens vers d'autres URI qui permettent de découvrir d'autres éléments.

L'idée consiste à appliquer l'architecture du World Wide Web [93] pour partager des données structurées à une échelle globale. Afin de comprendre ces principes, il est au préalable nécessaire de comprendre l'architecture du document Web classique. Le Web est construit sur un ensemble de standards simples :

- Des URI (Uniform Resource Identifiers, identifiants de ressource uniformes) comme mécanisme d'identification unique et global.
- HTTP (HyperText Transfer Protocol, protocole de transfert hypertexte), le mécanisme d'accès universel ;

- HTML (HyperText Markup Language, langage de balisage hypertexte), le format de contenu largement utilise.

De plus, il s'appuie sur le principe de liens existant entre des documents pouvant résider sur des serveurs différents. Les hyperliens permettent aux internautes de naviguer entre les serveurs et aux moteurs de recherche d'explorer le Web afin de fournir des fonctionnalités sophistiquées à partir du contenu explore. C'est pourquoi ces hyperliens sont centraux dans la connexion du contenu de différents serveurs afin de créer un espace d'information unique et global. En combinant la simplicité, la décentralisation et l'ouverture, le Web semble atteindre l'architecture idéale. Les données liées se fondent sur l'architecture du Web et l'appliquent au partage de données à l'échelle globale.

2.5.2 Nommer des éléments avec des URI

Pour publier des données sur le Web, il faut d'abord identifier les éléments du domaine d'intérêt. Il s'agit des éléments dont les propriétés et les relations seront décrites dans les données ; il peut s'agir de documents web, d'entités réelles et de concepts abstraits. Puisque les données liées s'appuient directement sur l'architecture du Web. Le terme ressource est réutilisé pour nommer ces éléments dignes d'intérêt, qui sont, à leur tour, identifiés par des URI HTTP. La Figure2.8 montre l'utilisation d'URI HTTP pour identifier des entités réelles et leurs relations. Sur cette photo de l'équipe de tournage de Big Lynx au travail, on voit le cameraman principal, Matt Briggs, avec ses deux assistants, Linda Meyer et Scott Miller, identifies par des URI HTTP de l'espace de noms Big Lynx. La relation (ils se connaissent) est représentée par des lignes, avec une URI de type http ://xmlns.com/foaf/0.1/knows. Les URI sont utilisées pour identifier des gens et les relations qui les joignent. Ces données liées n'utilisent donc que des URI HTTP (on évite les autres schémas d'URI, comme URN et DOI) et cela pour deux raisons :

- Les URI HTTP fournissent une manière simple de créer des noms globalement uniques, de façon décentralisée, puisque n'importe qui possédant un nom de domaine peut créer ou déléguer la création de références URI.

- Elles servent de nom mais aussi de moyen d'accès à l'information décrivant l'entité identifiée.

FIGURE 2.8 – URI sont utilisés pour identifier les personnes et les relations entre eux

2.5.3 Rendre les URI déréférençables

Toute URI HTTP devrait être déréférençables, autrement dit les clients HTTP devraient pouvoir visiter l'URI conformément au protocole HTTP et récupérer une description de la ressource identifiée. Cela s'applique aux URI nécessaires pour identifier des documents HTML classiques autant qu'à celles utilisées dans un contexte de données liées pour identifier des objets du monde réel et des concepts abstraits. Les descriptions des ressources sont incarnées par des documents web. Celles qui sont prévues pour être lues par des êtres humains sont généralement représentées en HTML et celles qui sont prévues pour être consommées par des machines le sont par des données RDF. Lorsque les URI identifient des objets réels, il est primordial de ne pas confondre les objets eux-mêmes avec les documents web qui les décrivent. Pour cela et afin d'éviter toute ambiguïté, on utilise couramment des URI différentes pour identifier l'objet réel et le document qui le décrit. Cette pratique permet d'effectuer des déclarations séparées sur un objet et sur le document qui le décrit. Par exemple, la date de création d'une personne peut être différente de celle du document qui la décrit.

Le Web est conçu comme un espace de données utilisables tant par les humains que par des machines. Les deux parties devraient être capables de récupérer des représentations des ressources dans une forme qui répond à leurs besoins. Cela peut être réalise avec le mécanisme HTTP appelé la négociation du contenu.

L'idée sous-jacente suppose que les clients HTTP envoient des en-têtes HTTP avec chaque requête pour indiquer les types de documents qu'ils préfèrent. Les serveurs inspectent ces en-têtes et répondent de façon appropriée : si l'en-tête indique que le client préfère le HTML, le serveur lui envoie un document HTML

et s'il préfère le RDF, le serveur lui envoie un document RDF.

2.6 Linked Open Data

Les données dans le Web sont reliées pour abaisser les obstacles d'accès à ces données. Le Mouvement Open Data [95] vise à rendre les données librement accessibles à tout le monde. Il existe déjà divers ensembles de données ouvertes intéressants disponibles sur le Web. Les exemples incluent Wikipédia, Wikilivres, Geonames, MusicBrainz, WordNet, etc.

Le but du Linked Open Data est d'étendre le Web avec des données communes en publiant divers ensembles de données ouvertes comme RDF sur le Web et par la mise en liens RDF entre éléments de données provenant de différentes sources de données.

2.7 Les types de données

Effectuer ces liens s'avère complexe car les données actuellement mises à disposition dans le Web de données sont la conversion de silos de données existants qui, par nature, ne sont pas reliés à d'éventuels autres ensembles de données. Cette difficulté a deux conséquences :

la nature des liens n'est pas très riche pour le moment, se contentant dans une majorité des cas d'indiquer une équivalence d'identité entre deux ressources avec la propriété owl :sameAs définie dans le vocabulaire OWL ; l'assouplissement des règles remplacées par un système de gradation croissante qui va de la mise à disposition des données sur le Web quel que soit le format au respect des quatre principes du Web de données.

De nombreux ensembles des domaines existent dans le cloud de LOD aussi variés que la musique, la géographie, le-gouvernement, la chimie. Le diagramme précédent fait apparaître sous des couleurs différentes les différents types de données présentes actuellement dans le Web de données qui en montrent la diversité et la richesse pour une initiative relativement récente, mais aussi les domaines encore absents.

2.8 Les bénéfices des données ouvertes liées

En ouvrant les données liées à tous, les données peuvent être réutilisées, partagées dans la communauté, ce qui crée un effet démultiplicateur. Les données libérées

et réutilisées génèrent ainsi des bénéfices à la fois dans les sphères économiques, culturelles, scientifiques, sociales, et technologiques. Nous essayons de résumer la plupart de ces bénéfices :

Publics et sociaux

En Grande-Bretagne, le croisement des données ouvertes a permis une diminution de 30% de la facture de consommation d'énergie dans les bâtiments publics en seulement deux mois. De plus, des voix s'élèvent pour rendre transparents les échanges économiques et financiers pour :

- éviter les problèmes systémiques qui plongent de nombreux pays dans une crise économique permanente.
- éviter les délocalisations abusives d'entreprises qui maquillent leurs comptes pour justifier leurs plans sociaux

Cependant, l'ouverture des données est souvent uniquement associée à l'idée de réduction des coûts. Les innovations de rupture par l'émergence d'usages inattendus sur ces données vont permettre, à titre d'illustration, la publication de données ouvertes sur les hôpitaux, permettant d'améliorer la qualité de leurs services.

Economiques

Les données ouvertes permettent, dans l'idéal, une concurrence équitable entre toutes les entreprises. Cependant, des études sociologiques en Inde et au Canada ont mis en évidence que l'accès et l'utilisation des données ouvertes étaient conditionnés par des critères matériels (électricité, possession de matériel informatique) et sociaux (éducation). De plus selon certaines études, la libération de ces données publiques diviserait par cinq le capital nécessaire pour exercer une activité professionnelle dans le secteur de la téléphonie mobile. Le rapport MEPSIR datant de 2006, financé par la commission européenne, estime que le marché européen lié à la réutilisation des informations publiques représente 27 milliards d'euros. Par ailleurs, l'impact économique direct et indirect a été évalué 140 milliards d'euros par an pour l'Europe

Culturelle et scientifique

L'ouverture des données scientifiques et le libre accès sont deux sujets connexes mais distincts. Le libre accès concerne les publications scientifiques, souvent relues par des pairs. L'ouverture des données scientifiques peut concerner les données à la base de ces articles, ou toute base de données à caractère scientifique (par exemple des relevés météorologiques ou autres), afin de permettre la reproduction des expériences menées, afin de les affirmer ou les infirmer.

2.9 Conclusion

Le Web de données, loin de revoir ses ambitions à la baisse, le Web sémantique en revient à ses origines. Certes, éloigné des promesses de l'intelligence artificielle, le Web de données propose un cadre d'interopérabilité pour mettre à disposition, lier et échanger des données structurées en vue d'un traitement simplifié par les machines. Or, cet enjeu est de taille, puisqu'il s'agit ni plus ni moins que de décloisonner les silos de données afin de libérer les usages faits de ces données. Il constitue donc une première étape indispensable pour envisager des traitements plus complexes sur cette masse de données.

Chapitre 3

Les Systèmes de recommandation à base de LOD

3.1 Introduction

Plusieurs approches ont été proposées pour incorporer le techniques du Web sémantique au système de recommandation, plus précisément les techniques de données ouvertes liées LOD. Dans ce chapitre, nous allons dresser un état de l'art des différentes approches des systèmes de recommandation fondées sur LOD. Cette revue de littérature est couronnée par un tableau comparatif faisant ressortir les différents points forts et faibles de chaque approche pour permettre de mieux positionner notre contribution.

3.2 [Heitmann et al ,2010]

3.2.1 Objectifs

L'objectif principale de [Heitmann et al, 2010] est d'établir un système de recommandation musical ouvert en utilisant les données ouvertes liées(LOD) afin d'atténuer les problèmes de système de recommandation collaboratif (les nouveaux utilisateurs, les nouveaux articles, la sparsity).

3.2.2 Principe de fonctionnement

La Figure 3.1 montre toutes les étapes de traitement des données liées pour des recommandations collaboratives. Les principales étapes sont :

1. Collecter les données de différentes sources (myspace, wikepedia) sous forme de triplet RDF. Myspace emploie :

- *Foaf :Person* classe pour les utilisateurs.
- *mo :MusicalArtist* pour les musiciens
- *myspace :topFriend* pour relier chacun des deux. et Wikepedia emploie :
  - *Foaf :Document* indique la page d'informations de l'utilisateur.
  - *sioc :links_ to* décrit un lien de cette page d'utilisateur
  - *sioct :WikiArticle* est un article de wikipedia dont le sujet intéressant à l'utilisateur

2. Transformer les données collectées en une représentation unifiée (*Foaf :Person, foaf :Interest, foaf :Document*) en utilisant une requête CONSTRUCT de SPARQL afin de demander à l'interface de données de récupérer les données et puis de fournir une règle pour exprimer les données à l'aide d'un vocabulaire différent

3. Transformer la représentation des données d'un graphe RDF en une matrice où les lignes représentent les usagers (foaf :Person), les colonnes représentent les items (*foaf :Document*) et leur connexion (*foaf :Interest*) est indiquée par une valeur binaire afin d'utiliser un filtrage collaboratif

4. Calculer la similarité entre deux items en utilisant la fonction *cosinus* suivante :

$$\cos(i_1, i_2) = count(i_1, i_2)/count(i_1), count(i_2) \tag{3.1}$$

   - $count(i_1)$, est le nombre d'usagers qui ont un lien avec l'item i_1
   - $count(i_1, i_2)$ est le nombre d'usagers qui ont un lien avec l'item i_1 et l'item i_2.
   - Classement des entrées dans la ligne i_1 fournit les éléments les plus similaires à i_2

3.2.3 Points forts

- Améliorer la précision et le rappel.
- Réduire les inconvénients du filtrage collaboratif (le démarrage à froid, le nouvel usager, le nouvel Item).
- Abaisser les barrières à l'entrée de l'exploitation d'un système de recommandation.

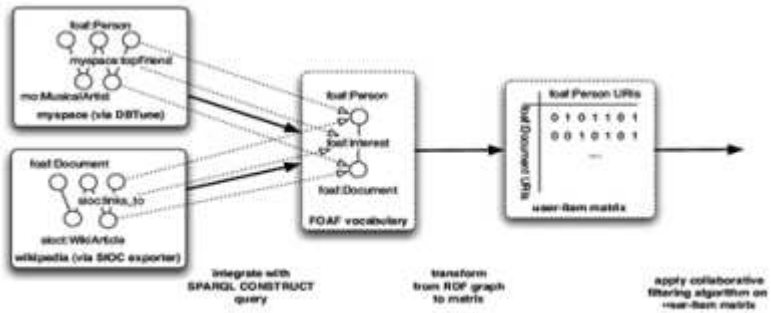

FIGURE 3.1 – Traitement de données liées pour les recommandations collaboratives

3.2.4 Points faibles

Malgré l'utilisation des LOD, le problème de sparsity persiste, vu que [Heitmann et al ,2010] a utilisé une matrice contenant tous les chanteurs et les chansons de son système pour applique le filtrage collaboratif

3.3 [Ostuni et al, 2012]

3.3.1 Objectifs

Comme [Heitmann et al ,2010], [Ostuni et al, 2012] ont créé un système de recommandation de film basé sur le contenu, en exploitant exclusivement des ensembles de données de LOD.

3.3.2 Principe de fonctionnement

Le principe général, consiste à représenter l'information par des vecteurs pondérés des mots-clés.

- Les items (films) du système sont enrichis par les données ouvertes liées (DBpedia, Freebase et LinkedMDB) en utilisant les triples RDF par SPARQL endpoints
- Chaque film est modélisé par VSM comme un vecteur pondéré en utilisant TF-IDF
- Le profil de l'utilisateur est modélisé par une notation binaire, j'aime/je n'aime pas

- Des algorithmes génétiques sont appliqués sur la base des vecteurs afin de réduire le sur-apprentissage.
- Des formules de similarité sont appliquées pour la recommandation.

L'approche utilise un VSM (Vector Space Model) sémantique pour représenter l'ensemble graphe RDF comme une matrice de 3-dimensions où chaque tranche se réfère à une propriété de l'ontologie et représente sa matrice d'adjacence (Figure 3.2). Les tranches de la matrice sont décomposées en plusieurs petites matrices où chaque matrice se réfère à une propriété RDF spécifique où les lignes représentent le domaine de la propriété, les colonnes sont le co-domaine (rang). Un film m est alors

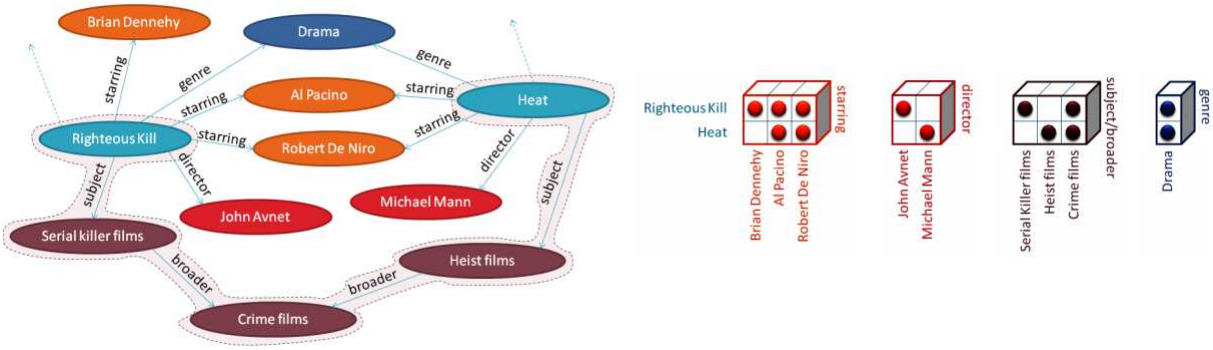

FIGURE 3.2 – La représentation matricielle d'un graphe RDF de domaine de film.

représenté par un vecteur contenant tous les termes/noeuds(n) liés à m via p (le nombre total de propriétés sélectionnées). Le système évalue le degré de similitude entre le film m_i et film m_j par rapport à p, Par la corrélation entre les vecteurs $w_{n,i,p}$ et $w_{n,j,p}$ Plus précisément le cosinus de l'angle entre les deux vecteurs se calcule comme suite :

$$sim(m_i, m_j) = \frac{\sum_{n=1}^{t} w_{n,i,p} \times w_{n,j,p}}{\sqrt{\sum_{n=1}^{t} w_{n,i,p}^2} \times \sqrt{\sum_{n=1}^{t} w_{n,j,p}^2}} \tag{3.2}$$

Le profil d'utilisateur repose sur une notation binaire (j'aime / je n'aime pas)

$$profile(u) = h_{mi}, v_{ji} | v_j = 1 \; if \; u \; likes \; m_j, v_j = -1 \; otherwise \tag{3.3}$$

Afin d'évaluer si une nouvelle ressource mi (film) pourrait être d'intérêt pour un utilisateur u, le système combine les valeurs de similarité liées à chaque propriété unique des mi et calcule une valeur de similitude globale r (u, mi). Deux formules sont proposées :

$$profile(u) = h_{mi}, v_{ji} | v_j = 1 \; if \; u \; likes \; m_j, v_j = -1 \; otherwise \tag{3.4}$$

Afin d'évaluer si une nouvelle ressource m_i (film) pourrait être d'intérêt pour un utilisateur u, le système combine les valeurs de similarité liées à chaque propriété unique des m_i et calcule une valeur de similitude globale $r(u, m_i)$. Deux formules sont proposées :

$$\overline{r}(u, m_i) = \frac{\sum_{m_j \in profile(u)} v_j \times \frac{\sum_p \alpha_p \times sim^p(m_j, m_i)}{p}}{|profile(u)|} \tag{3.5}$$

$$\overline{r}(u, m_i) = \frac{\sum_{m_j \in profile(u)} v_j \times \frac{\sum_p \alpha_p \times sim^p(m_j, m_i)}{p}}{|\sum_p \alpha_p|} \tag{3.6}$$

Le système calcule automatiquement les valeurs par défaut en formant le modèle via un algorithme génétique qui offre des fonctions de forme physique et minimise l'erreur de classification fausse sur les données de formation améliorant de ce fait les résultats de précision. Une autre approche de l'extraction automatique des poids, basés sur Amazon est utilisé lorsqu'un nouvel utilisateur commence à utiliser l'application.

3.3.3 Points forts

- Réduire le problème du nouvel item
- Amélioration du rappel et de prédiction.
- Réduire le sur-apprentissage.

3.3.4 Points faibles

La représentation matricielle de l'ensemble des films du système de recommandation de [Ostuni et al, 2012] a généré une grande sparsity

3.4 [Ostuni et al, 2013]

3.4.1 Objectifs

- Calculer les Top-N recommandations des articles de feedback implicite (film) en exploitant les informations disponibles dans le Web des données
- Extraire les caractéristiques basées sur les chemins sémantiques pour calculer par la suite des recommandations utilisant un algorithme d'apprentissage automatique (classement).

3.4.2 Principe de fonctionnement

L'approche de [Ostuni et al, 2013] suit les étapes suivantes :

1. Construire un graphe qui représente l'élément intéressant à utilisateur à partir de la matrice de feedback implicite.
   - Les utilisateurs et les articles sont les nœuds et les feedback des utilisateurs sont les liens.
   - $I_u^+ = i \in I|\widehat{s}_u i = 1$ L'ensemble des items pertinents pour $U.$, i.e les items qui sont reliés avec l'utilisateur par un arc (Figure 3.3)
   - $I_u^- = i \in I|\widehat{s}_u i = 0$ L'ensemble des items non pertinents pour $U.$., i.e les items qui ne sont pas reliés avec l'utilisateur par un arc (Figure 3.3)
   - $I^- \times u \subseteq I_u^-$ échantillon des items non pertinents pour U.
2. Utilisation des données sémantiques structurelles librement disponibles sur le Web (Dbpedia) pour décrire les items dans un graphe.
3. Combinaison des descriptions d'article sémantique à partir du Web des données (étape 2) et des feedbacks implicites (étape 1) pour les tâches de Top-N recommandation (Figure 3.3).
   - U, I et E représentent respectivement les utilisateurs, les articles et les entités
   - un nouveau graphe $G = (V; R)$ où V dénote l'ensemble de sommets et de R l'ensemble de relations.
   - R contient deux types de relations $S = U \times I$, $P \subseteq E \times E$ et$R = S \cup P$.
4. Analyse de relation complexe entre les préférences d'utilisateurs et la cible d'item (extraction des chemins).

Codons les caractéristiques capables de caractériser l'interaction entre l'utilisateur u et l'item i dans le vecteur Xui RD où D est la dimension de l'espace de caractéristique.

$$X_{ui}(j) = \frac{\#path_{ui}(j)}{\sum_{d=1}^{D} \#path_{ui}(d)} \tag{3.7}$$

- $\#path_{ui}(j)$ nombre de chemins entre u et i

5. construire l'ensemble de formation (training set)

$$\bigcup_u \langle x_{ui}, \widehat{s}_{ui} \rangle | i \in (I^+, I^{-*}) \tag{3.8}$$

- $\widehat{s}_u i$ c'est la matrice binaire de feedback implicite

6. Formalisation du problème de Top-N recommandation d'article de feedback implicite dans un apprentissage automatique de classement
   - Fonction de classement $f : R^d$ sachons que $f(X_{ui})\widehat{s}_{ui}$
   - Hypothèse : si f est précis, alors le rang induit par f devrait être proche du rang désirant.
   - Utiliser l'algorithme de classement bagboo qui fusionne deux autres algorithmes Random Forests and Gradient Boosted Regression Trees

3.4.3 Les points forts

- Améliorations significatives par rapport aux méthodes de filtrage collaboratif en particulier en cas de forte sparsity.
- Améliorer le rappel.
- Calculer des recommandations précises.
- Réduire la variance d'une fonction de prédiction estimée.
- Réduire le sur-apprentissage.
- Réduire les inconvénients du filtrage collaboratif (démarrage à froid).
- Représentation uniforme pour le filtrage collaboratif et à base de contenu dans un seul graphe

3.4.4 Points faibles

[Ostuni et al, 2013] a relié l'utilisateur avec un item dans son graphe sans prendre en considération la valeur de l'évaluation (rating) i.e même si l'utilisateur U a donné une évaluation de 1 ou 2 à item I. Le système [Ostuni et al, 2013] peut recommander à cet utilisateur U des items qui sont similaire à l'item I

3.5 [Yang et al, 2013]

3.5.1 Objectifs

- Construire un système de recommandation des films basé sur l'algorithme Slope One en utilisant les technologies sémantiques.

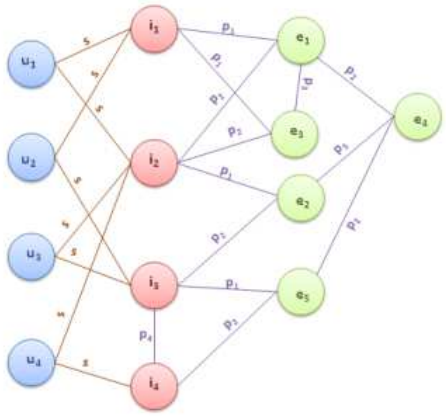

FIGURE 3.3 – Combinaison des graphes (étape 1 et 2)

- Combiner les technologies sémantiques avec le filtrage collaboratif traditionnel pour augmenter le nombre de précision sans réduire l'efficacité de calcul et la simplicité
- Explorer les relations implicites entre les éléments basés sur les données liées et des mesures pour le calcul des distances sémantiques

3.5.2 Le principe de fonctionnement

La Figure 3.4 montre les principales étapes de l'approche proposée par [Yang et al, 2013]

1. Identifiez le sous-ensemble (dbTropes) approprié de LOD (sous forme de triples de RDF).
2. Mapper les URIs(titres des films) en triples RDF aux items dans des ensembles de données traditionnelles(MoviLens)(trouver les correspondances entre les titres des deux sources en utilisant la correspondance par chaîne de caractères).
3. Employez l'algorithme de *LDSD* pour calculer les distances sémantiques et pour les insérer dans un ensemble de données traditionnelles comme similitudes d'item-à-item.

$$LDSL_{d(r_a,r_b)} = \frac{1}{1 + C_d(n, r_a, r_b) + C_d(n, r_b, r_a)} \tag{3.9}$$

$$LDSL_{i(r_a,r_b)} = \frac{1}{1 + C_i o(n, r_a, r_b) + C_i i(n, r_b, r_a)} \tag{3.10}$$

- r_a, r_b deux ressources, n le lien entre ces deux ressources.
- $LDSD_{d(r_a,r_b)}$: elle considère seulement les liens directs entrants et sortants.
- $LDSD_{i(r_a,r_b)}$: elle prend les liens indirects en considération.
- $C_{d(n,r_a,r_b}$) est une fonction qui calcule le nombre de liens directs et distincts entre les ressources dans un graphe G. égale 1 s'il y a une instance de n de la ressource ra à la ressource rb
- C_{io} etC_{ii} sont deux fonctions qui calculent le nombre d'indirects et distincts liens sortants et entrants, entre les ressources dans un graphe G.
- $C_{io(n,r_a,r_b)}$ égale 1 s'il y a une ressource k qui satisfait chacun des deux $\{n, r_a, k\}$ et $\{n, r_b, k\}$, sinon 0.
- $C_{ii(n,r_a,r_b)}$ égale 1 s'il y a une ressource k qui satisfait chacun des deux $\{n; k; r_a\}$ et $\{n; k; r_b\}$, sinon 0

4. Intégrer les similitudes dans le poids de Slope One original et calculer les recommandations.

5. équation Non-Linéaire : L'estimation de prévision non linéaire de l'usager u à l'item i est calculée comme suit :

$$P_u i = \frac{\sum_{i \in I_u - i_j} (\sum_{x \in S_{ji}} \frac{v_{xi} - v_{xj}}{|S_{ji}|} + v_{ui}) log \frac{|S_{ij}|}{LDSD(i_i, i_j)}}{\sum_{i \in I_u - i_j} log \frac{|S_{ij}|}{LDSD(i_i, i_j)}} \tag{3.11}$$

6. Elle change directement la méthode de calcul de poids dans l'équation original. équation Linéaire :

$$P_{ui} = (1 - \alpha) \times \frac{\sum_{i \in I_u - i_j} (\sum_{x \in S_{ji}} \frac{v_{xi} - v_{xj}}{|S_{ji}|} + v_{ui})}{\sum_{i \in I_u - i_j} |S_{ji}|} + A \tag{3.12}$$

$$A = \alpha \times \frac{\sum_{i \in I_u - i_j}) \sum_{x \in S_{ji}} (v_{xi} - v_{xj}) + v_{ui} |S_{ji}|) \frac{1}{LDSD(i_i, i_j)}}{\sum_{i \in I_u - i_j} \frac{1}{LDSD(i_i, i_j)}} \tag{3.13}$$

La première moitié de l'équation est l'équation originale Slope One, et la dernière moitié prend en considération le LDSD au lieu du nombre d'estimations fournies par chaque utilisateur. Le paramètre gratuit (α) est employé pour ajuster la proportion.

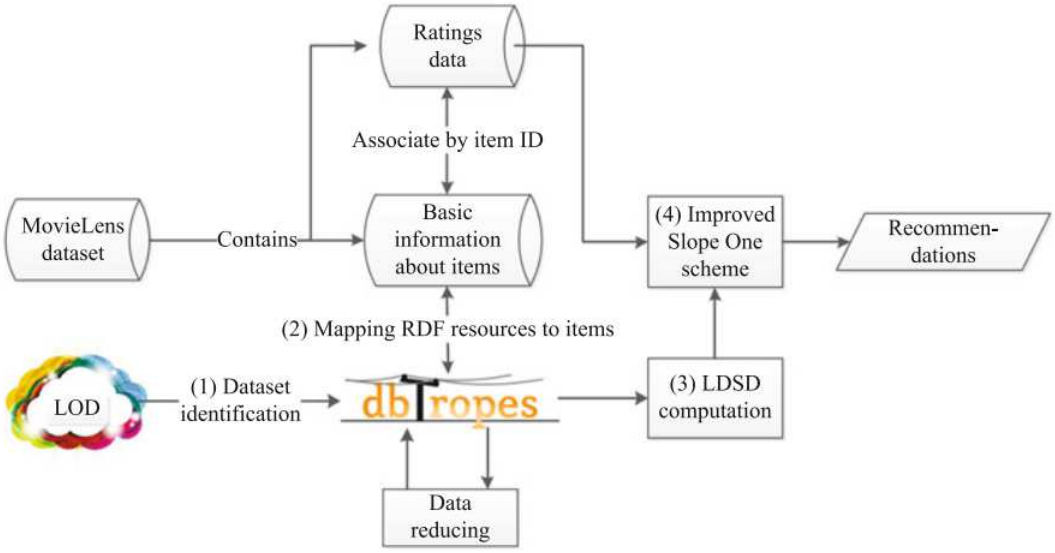

FIGURE 3.4 – Vue d'ensemble de l'approche

3.5.3 Les points forts

- Rendre la prévision d'estimation plus précise.
- Les liens directs donnent des relations plus fermes.
- $LDSD_d$ contient l'information efficace.
- $LDSD_d$ trouve les relations implicites correctes.
- Effet positif de $LDSD_d$ quand le paramètre gratuit varie.
- Augmenter la précision et le rappel.

3.5.4 Les points faibles

- $LDSD_i$ liens indirects peuvent provoquer quelques doutes.
- Non amélioration de précision dans la méthode non linéaire.
- La couverture non augmenté (problème de sparsity)

3.6 [PESKA et al,2013]

3.6.1 Objectifs

Amélioration des systèmes de recommandation des livres d'occasion en utilisant les Linked Open data par :

- Proposer une méthode en ligne pour enrichir automatiquement les attributs d'un objet avec linked open data afin d'améliorer la recommandation basée sur le contenu.
- évaluation expérimentale avec les deux types de filtrages (collaboratifs et basés sur le contenu

3.6.2 Le principe de fonctionnement

Le schéma suivant illustre les différentes étapes de l'approche :

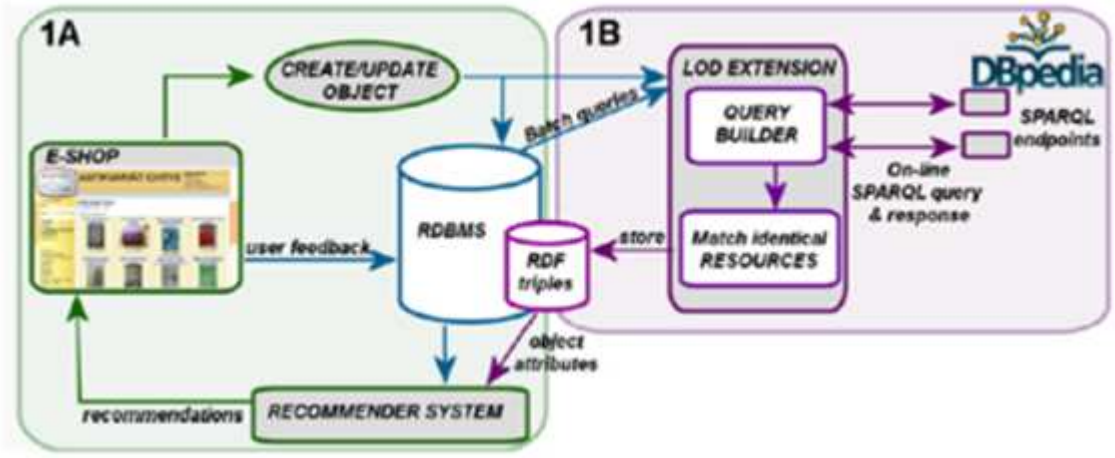

FIGURE 3.5 – L'architecture de l'amélioration du système de recommandation avec LOD.

1. Le système maintient un ou plusieurs SPARQL endpoints à divers ensembles de LOD (datasets). Les connexions sont habituellement des REST APIs ou des services simples de HTTP.
2. Chaque fois qu'un objet du système est créé ou mis a jour, le système automatiquement établi une requête avec chacune des connexions de SPARQL avec l'identificateur unique d'objet ou tout autres meilleures spécifications d'objet.

3. Les triples RDF sont stockés dans le magasin de triples RDF du système (ex : base de données relationnelle).

4. Les triplets RDF sont Transformer en une relation binaire < object, attribute>.

5. La construire d'une matrice booléenne à partir de ces relations binaires.

6. Le système devrait interroger régulièrement les datasets (LOD) pour chaque objet (en gardant une copie locale de LOD datasets). Chaque objet du système est représenté par son vecteur d'attributs (numérique, String, ensemble, non-ordonner).

   - **Attributs numériques** :Normalisation des valeurs absolues et des attributs numérique
   $$|a_x - a_y|/max_{\forall objectso}(|a_0) \tag{3.14}$$
   - **Attributs de type String** : levenshtein distance
   $$1 - (levenshtein(a_x, a_y)max(lenght(a_x), lenght(a_y))) \tag{3.15}$$
   - **Attributs non-ordonner** : L'égalité des attributs.
   - **Attributs de type ensemble** :la similarité de Jaccard
   $$(a_x \bigcap a_y)/(a_x \bigcup a_y) \tag{3.16}$$

3.6.3 Points forts

- Réduire le problème de sparsity des données.
- Abaisser le problème du nouvel item.

3.6.4 Points faible

- Charge excessive à la fois des données de stockage et de maintenance du système.
- Les problèmes du démarrage à froid et de nouvel usager résident

3.7 [Ku et al, 2014]

3.7.1 Objectifs

L'approche novatrice propose une recommandation basée sur le contenu au sein de différentes catégories multimédia (film, série, etc.) en tenant compte à la fois l'information sémantique de contenu et les intérêts des utilisateurs.

3.7.2 Le principe de fonctionnement

La Figure 3.6 montre les différentes étapes de l'approche

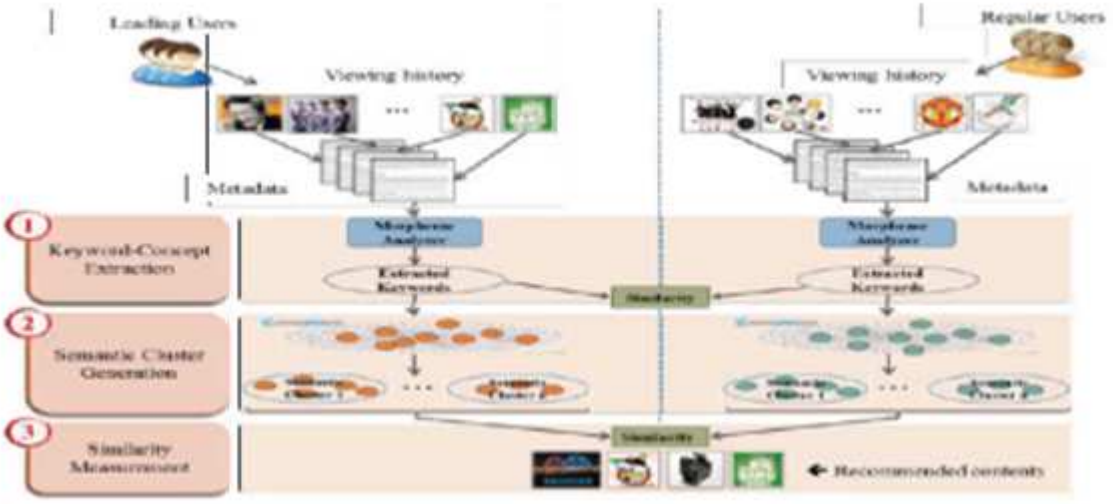

FIGURE 3.6 – Processus global de l'approche

1. Identifier premièrement les principaux utilisateurs qui ont consommé les divers contenus intensivement pendant une certaine période, en comptant principalement le nombre de différents contenus qu'ils ont consommés au cours d'une période.

2. Pour chaque utilisateur principal, le système produit un groupe sémantique qui représente sa préférence en consommant le contenu.

3. Le système accède aux données liées pour obtenir les concepts sémantiques qui sont liés au contenu que les principaux utilisateurs ont consommé.

4. Le système les groupe dans des ensembles basés sur la similitude entre groupes sémantiques. Le système appelle ces groupes en tant que principaux groupes d'utilisateurs (leading user groups).

5. Pour chaque utilisateur général (non-principal utilisateur), ils ont produit également des groupes sémantiques basés sur le contenu qu'ils ont consommé.

6. Le système classifie les utilisateurs généraux sous les groupes de principaux utilisateurs basés sur la similitude entre les groupes sémantiques des principaux groupes d'utilisateurs et ceux des utilisateurs généraux. Un utilisateur général peut être classifié sous un multiple groupe d'utilisateur principal.

7. Ils constatent la différence entre les contenus qui sont consommés par un groupe principal d'utilisateurs et le contenu qui est consommé par chacun des utilisateurs généraux classifiés sous le groupe.

8. Une liste de contenu qui est consommé par le groupe principal d'utilisateurs, mais n'ont pas consommé par l'utilisateur général dans le groupe sera recommandé à l'utilisateur général comme contenu potentiellement intéressant. La génération des clusters sémantiques des utilisateurs se fait selon les étapes suivantes :

    - Récupérer les concepts qui sont liés aux mots-clés trouvés dans le contenu consommés par l'utilisateur on utilisant linked open data pour les identifier.
    - L'extraction des noms et des groupes nominaux à partir des descriptions de contenu à partir des linked open data.
    - Traité plusieurs prédicats tels que rdf :label et skos :prefLabel utilisés dans les datasets.
    - établissement d'une nouvelle requête pour récupérer tous les triples RDF qui décrivent chaque sujet et de les regrouper dans un concept.
    - Construire des groupes pertinents, appelés clusters sémantiques pour filtrer les concepts non relatifs.

- *hasKeyword* : relie le concept par le label de la présentation.
- *hadURI* : relie le concept avec sa référence URI.
- *owl :sameAs, skos :exactMatch* : consolide les concepts qui ont le même sens.
- *skos :broader , skos :narrower* : trouvons et fusionnons sémantiquement des concepts appropriés.

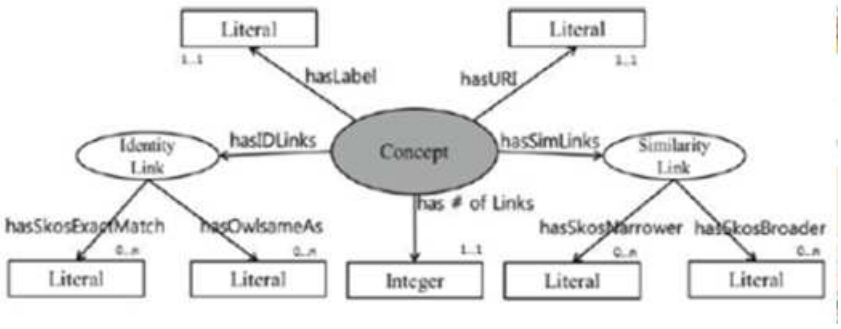

FIGURE 3.7 – Processus global de l'approche

La Mesure de la similitude entre les utilisateurs est calculée comme suit :

$$sim(lc, rc) = \sum C(K_{lc-rc}) \tag{3.17}$$

- Lc : le cluster sémantique généré pour l'ensemble de contenu des utilisateurs principaux.
- rc : le cluster sémantique généré pour l'ensemble de contenu des utilisateurs généraux.
- k_{lc-rc} : tous les concepts qui sont coproduits dans les deux ensembles.

3.7.3 Points forts

- Réduire le problème de sparsity des données.
- éliminer le problème de recommandation collaborative dans le domaine de différentes catégories.
- Réduire le problème de nouveaux usagers

3.7.4 Points faibles

Le système de [Ku et al, 2014] perd de la précision lors de la recommandation à un utilisateur atypique car c'est un utilisateur qui n'est pas similaire à aucun groupe d'utilisateurs.

3.8 [Ragone and al, 2017]

3.8.1 Objectifs

[Ragone and al, 2017] utilisent un schéma récapitulatif nommé ABSTAT. ABSTAT prend l'ensemble des linked data et renvoie un Sommaire. Ce Sommaire est un motif représenté sous forme de m <C, P, D>, où C et D sont les concepts et les types de données, et P representé les propriétés RDF. Chaque motif indique qu'il existe une instance de type C lié à une instance de type D à travers la propriété P. ABSTAT est transformé en deux graphes. Le premier est basé sur Entity-based item neighborhood mapping et le deuxieme Path-based item neighborhood mapping

3.8.2 Le principe de fonctionnement

Dans cette méthode le mapping est caractérisé par une entité E et son poids. Le vecteur résultant est :

$$_E(G_i^h) = (w_{i,e_1}, w_{i,e_2}, ..., w_{i,e_m}, ..., w_{i,e_t}) \tag{3.18}$$

Où le poid associé à l'entité e_m est calculé par :

$$w_{i,e_m} = \sum_{h}^{l=1} \alpha_l . c_{l,e_m} \ with \alpha_l = \frac{1}{1 + log(l)} and \ A \tag{3.19}$$

$$A = c_{l,e_m} = |(e_n, p, e_m)| e_n \in \hat{E}_i^{l-1} \wedge e_m \in \hat{E}_i^l \bigcup B \tag{3.20}$$

$$B = (e_n, p, e_m) | e_m \in \hat{E}_i^l \wedge e_n \in \hat{E}_i^{l-1} | \tag{3.21}$$

où $\hat{E}_i^l = \hat{E}_i^l \ \hat{E}_i^{l-1}$ est l'ensemble des dataset
Contrairement au mappage précédent, une caractéristique est représentée comme une séquence de nœuds dans G. Chaque caractéristique se réfère à plusieurs variantes de nœuds enracinés dans le nœud i. Ces nœuds sont collectés. Ensuite, à partir de ces chemins, d'autres fonctionnalités sont définies en fonction de chaque sous-chemin. Plus précisément, les sous-chemins sont composés uniquement de ces entités progressivement plus éloignées de l'objet. Cette représentation permet de représenter explicitement les sous-structures partagées entre les éléments sans chevauchement dans leurs quartiers immédiats mais en quelque sorte connectés à plus grande distance.

3.9 Synthèse

Après avoir étudié les différentes approches de systèmes de recommandations fondées sur LOD. Nous avons réalisé un tableau comparatif ci-dessous entre ces différentes approches. Le tableau4.1 montre une comparaison entre les approches effectuées selon sept critères (domaine, Sources LD, Feedback, Filtrage, techniques et Algorithmes, Formule de similarité, représentation de données). On peut classer les approches étudies dans l'état de l'art dans trois catégories selon les types de filtrage

3.9.1 Filtrage Collaboratif

- **Basé mémoire**

  [Heitmann et al, 2010] ont proposé un système de recommandation de musique, ils ont collectés des données pour enrichir les items et les utilisateurs de différentes sources LOD (myspace, wikepedia) sous forme de triplet RDF et les transformer en une représentation RDF unifiée puis une matrice binaire pour appliquer le filtrage collaboratif. [Yang et al, 2013] ont rassemblé des données à partir DBTropes comme LD et comme source de données traditionnelles MovieLens en utilisant le feedback explicite, et puis faire la correspondance entre les titres de ces deux sources de données, ensuite, ils ont Calculé la prédiction de l'usager en utilisant l'algorithme LSLD, enfin faire la recommandation en utilisant l'algorithme Slope one.

- **Basé modèle**

  [Ku et al, 2014] proposent une recommandation de différentes catégories multimédia (film, série, etc.) en tenant compte à la fois l'information sémantique de contenu et les intérêts des utilisateurs, utilisant des groupes sémantiques générés pour chaque utilisateur, les LD ont été utilisés dans la production des clusters en recueillant plus d'informations sur les films vus par les utilisateurs. Cette approche utilise le feedback implicite.

3.9.2 Filtrage à base de contenu

[Ostuni et al, 2012] ont Créé un système de recommandation de film basé sur le contenu pour un feedback explicite (like/dislike), en exploitant exclusivement des ensembles de données de LOD (DBpedia, Freebase et LinkedMDB). Le principe du système de recommandation de l'approche est l'information représentée

par vecteurs pondérés des mots-clés (VSM Modèle vectoriel). Puis appliquer des algorithmes génétiques sur la base des vecteurs.

[Peska et al, 2013] ont amélioré les systèmes de recommandation basés sur le contenu des livres d'occasion en utilisant les linked Open data (dbpedia) où ils ont proposé une méthode en ligne pour enrichir automatiquement les attributs d'un objet avec linked open data. Ils utilisent le feedback implicite.

3.9.3 Filtrage Hybride

[Ostuni et al, 2013] offrir les meilleures N recommandations par rapport aux feedbacks implicites en utilisant les données liées (dbpedia), cette approche permet de recommander des items (films) en exploitant leurs propriétés et attributs qui sont définis dans un graphe sémantique. Enfin pour classer les meilleures recommandations en appliquant l'algorithme de classement bagboo

Approche	Domaine	Sources LOD	Feedback	filtrage	Technique et algorithme	Formule de similarité	Représentation des Données
[Heitmann et al, 2010]	Music	Myspace, DBpedia	Explicit	Collaboratif	Matrice Usager-item	Cosinus	FOAF
[Ostuni et al, 2012]	Film	DBpedia, LinkedMDB, Freebase	Explicit	À base de contenu	matrice de 3 dimensions, Vector Space Model (VSM)	Cosinus	Triplet RDF
[Ostuni et al, 2013]	Film	DBpedia	implicit	Hybride	Graphe, Baggboo(Random Forests (RF), Regression Trees(GBRT)	/	Triplet RDF
[Yang et al, 2013]	Film	DBTropes	Explicit	Collaboratif	Slone One	Cosinus ou jaccard LDSD	Triplet RDF
[Peska et al, 2013]	Livre d'occasion	DBpedia	Implicit	à base de contenu	Marice binaire	Cosinus	Triplet RDF <Object,Attribut>
[Ku et al, 2014]	multimédia	DBpedia, Freebase, Linked Movie Database	Implicit	Collaboratif	Clusters	count	Triplet RDF <Object,Attribut>
[Ragone and al, 2017]	Film	DBpedia	Implicit	à base de contenu	graph-based kernel methods	/	Triplet RDF

TABLE 3.1 – Tableau comparatif entre les différentes approches de système de recommandation basé sur les LOD

3.10 Conclusion

Cet état de l'art a présenté un certain nombre d'approches visant à produire des systèmes de recommandation en utilisant les données ouvertes liées. Les approches des systèmes de recommandation sont particulièrement variées, et peuvent être classées de différentes manières (collaboratif, à base de contenue, hybrides, etc.). Toutes ces approches présentent des caractéristiques complémentaires. Par conséquent de nombreux travaux se sont intéressés aux différentes techniques d'hybridation, qui s'avèrent fournir des recommandations plus précises, en outre ils permettent de profiter des avantages respectifs de ces approches.

Dans le chapitre suivant nous allons vous présenter notre approche qui tente à améliorer les systèmes de recommandations hybride de film (Collaboratif, à base de contenu) pour un feedback explicite, en exploitant exclusivement les données liées (dbpedia). Les films après l'enrichissement sont représentés par des vecteurs pondérés des mots-clés (VSM Modèle vectoriel) et les utilisateurs sont représentés par un graphe. Les utilisateurs, les items et leurs entités sont définis dans un graphe sémantique

Chapitre 4

Approche sémantique pour l'amélioration des RS

4.1 Introduction

Les systèmes de recommandation tentent d'exploiter le maximum des données disponibles dans le web pour répondre aux besoins des utilisateurs. Ces données sont en grande quantité, elles nécessitent un filtrage pour prendre les meilleures données. Actuellement, la plupart des systèmes de recommandation de film utilisent l'algorithme de filtrage collaboratif ou celui du filtrage à base de contenu. Ces deux types de filtrage souffrent de plusieurs problèmes tels que le démarrage à froid, le nouvel usager, nouvel item, etc. Dans ce travail de thèse, nous proposons une approche hybride qui combine le filtrage collaboratif avec le filtrage à base de contenu tout en collectant des données auprès LOD, et ce afin de minimiser les problèmes mentionnés préalablement.

4.2 Architecture générale du système

Dans cette section, nous allons décrire l'architecture globale de notre système basé sur les données ouvertes liées. Cette architecture est divisée en deux principales parties (voir Figure 4.1) :

1. **La préparation des données :**
   Cette étape, qui est exécutée en mode Offline, permet de :
   - Enrichir les items à partir de Dbpedia.
   - Construire un graphe qui relie entre les utilisateurs similaires.

- Construire un vecteur Space Model, afin de calculer les similitudes entre les films
- Construire un graphe sémantique : qui représente la liaison entre l'utilisateur et l'item et la liaison entre un item et un autre.

2. **La recommandation :**
   Cette étape, qui est exécutée en mode Online, permet :
   - La génération des chemins.
   - Le filtrage des chemins.

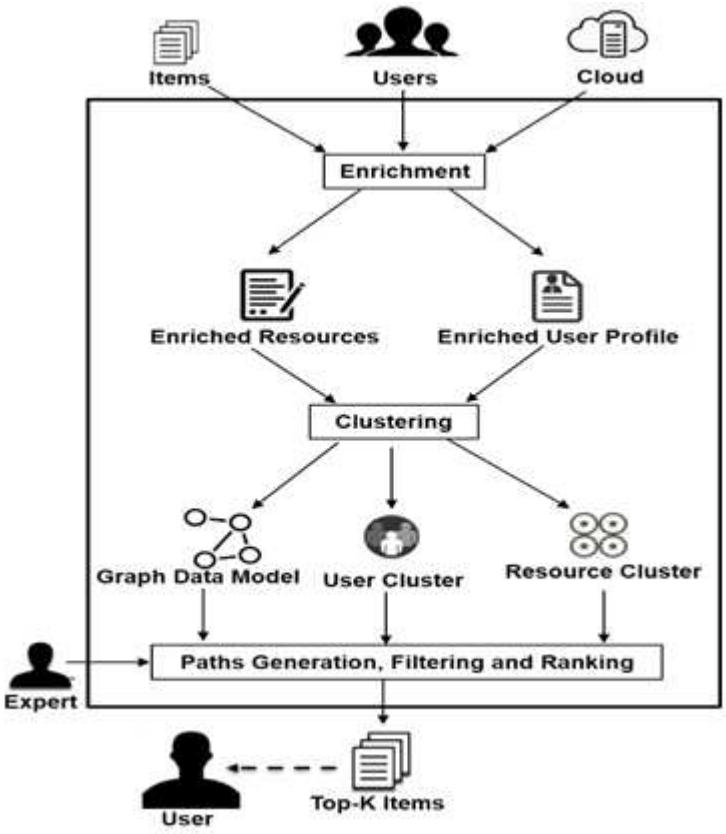

FIGURE 4.1 – Architecture générale du Système

4.3 Ressources

Notre système de recommandation est décrit par trois ressources (U, I, E) qui représentent l'utilisateur, l'item et l'entité.

4.3.1 Utilisateur

Les éléments U comprennent tous les utilisateurs qui ont vu l'item où contribué aux évaluations. L'utilisateur est la personne qui utilise le système de recommandation, il donne son avis sur une ressource et il reçoit les recommandations.

L'utilisateur est décrit par un profil qui définit ses centres d'intérêts. De tels profils consistent en un ensemble d'informations qui peuvent être entrées manuellement par l'utilisateur.

Chaque utilisateur U (profils utilisateurs) est représenté par *(Id, nom, sexe, âge, préférence =*$(Motscl_1$ *: valeur) ,* $(Motscls_2$ *: valeur),* $(Motscls_3$ *: valeur), etc.).*

Exemple :
U =(47, Amina, (Acteur : Jack Lemmon) , (Film : basketball), (Film : Grumpier Old Men), (Livre : l'alchimiste)).

4.3.2 Item

Les items sont produits dans le système de recommandation pour les suggérer à l'utilisateur. L'item I est représenté par *(id, nom, description =* $(Motscls_1$ *: valeur),* $(Motscls_2$ *: valeur), (motscls*$_3$*, valeur), etc.).*

Exemple :
I = (12, Jumanji, (année : 1995), $(Genre_1$ *: enfants),* $(Genre_2$ *: comedy),* $(Genre_3$ *:aventure))).*.

4.3.3 Entité

Les entités E sont les valeurs des propriétés P qui décrivent un item dans un système de recommandation. Comme il est cité au-dessous, l'item I est représenté par : *(id, nom, description =* $(P_1 : E_1)$*,* $(P_2 : E_2)$*,* (P_3, E_3)*, etc.).*

4.4 Principe de fonctionnement

Notre système de recommandation est un système qui est décomposé en deux principales parties : la préparation des données et la recommandation.

4.4.1 Préparation des données

La première partie de notre système de recommandation concerne la préparation des données, elle est divisé en quatre étapes : l'enrichissement, la construction d'un graphe d'utilisateur, la construction d'un Vector Space model(VSM), et la génération d'un modèle de donnée.

4.4.1.1 L'enrichissement

Un enrichissement sémantique est une information supplémentaire qui est ajoutée aux données de certaines ressources. Il a été perçue par les utilisateurs comme un moyen de réponse au problème de la qualité et de la complétude des données que l'on peut trouver.

Afin d'enrichir le profil d'un utilisateur ou un item, un système de recommandation peut bénéficier de la richesse du web des données. En particulier, le jeu de données DBpedia fournit des informations riches dans une variété de domaines. En utilisant les appariements entre les artistes de DBpedia et MovieLens, nous avons pu enrichir la description d'un film de MovieLens par ses acteurs et ses écrivains de DBpedia.

Notre système maintient DBpedia SPARQL Endpoints à divers ensembles de données LOD. Il récolte la description de chaque élément du système à partir de l'endpoint. À la fin de cette phase, la description des articles sera enrichie. Après l'enrichissement de I, il sera transformé en I', où $I' = I + E$.

La requête SPARQL qui permet l'interrogation de la base DBpedia est la suivante afin d'enrichir un film par la liste des acteurs est :

```
"PREFIX db : <http ://dbpedia.org/resource/>"
+ "SELECT distinct ?starring " + "WHERE " + "db :"
+ titre
+ " <http ://dbpedia.org/ontology/starring> ?starring"
+ "";
```

Exemple :

Soit I = (12, *Jumanji, (année : 1995), (Genre*$_1$ *: enfants), (Genre*$_2$ *: aventure), (Genre*$_3$ *: fantasy))).*

Après enrichissement, I devient $I' = (1_2$, *Jumanji, (année : 1995), (Genre*$_1$ *: enfants), (Genre*$_2$ *:aventure), (Genre*$_3$ *: fantaisie), (Acteur*$_1$ *: Robin Williams), (Acteur*$_2$ *: Kirsten Dunst Diesel), (Ecrivain*$_1$ *: Kirsten Dunst Diesel)))..*

A la fin de la phase d'enrichissement des items,Un Vector Space Model est construit.

Ce modèle est utilisé ulterieurement dans la phase de recommandation.

4.4.1.2 Clustering

Dans cette phase, les éléments seront regroupés en fonction de leurs similitudes sémantiques. Nous calculons la similitude entre les descriptions sémantiques des articles. De même, les utilisateurs seront également regroupés à l'aide de leurs profils et évaluations.

Dans ce travail de thèse, de nombreux algorithmes de cluster ont été appliqués aux ensembles de données. Ces algorithmes comprennent, mais sans s'y limiter, les K-means, la similarité du graphique et le VSM.

1. ***Utilisateurs***
   Pour le Clustering des utilisateurs, nous avons utilisé deux algorithmes : K-means et le graphe de similarité. Le but de K-means est de diviser les observations en K clusters dans lesquels chaque observation appartient à la grappe avec le moyen le plus proche (Voir Algorithme 01).

Algorithme01 : clustering des utilisateurs

```
Require : Cluster C, Number of Clusters k, Number of user NU.
Ensure : K clusters of Users
    for i=1 to NU do
        Compute Sim(i,j)
        C is Cluster(i)
        Cluster(j,C)
    end for
```

   Le graphe de similarité est un graphe qui relie les utilisateurs les uns avec les autres, en se fondant sur une mesure de similarité (figre4.2). Le graphe G = (U, S) où U représente les utilisateurs et S représente la mesure de similarité entre utilisateurs. Cette partie de l'approche est divisée en trois sous-parties :

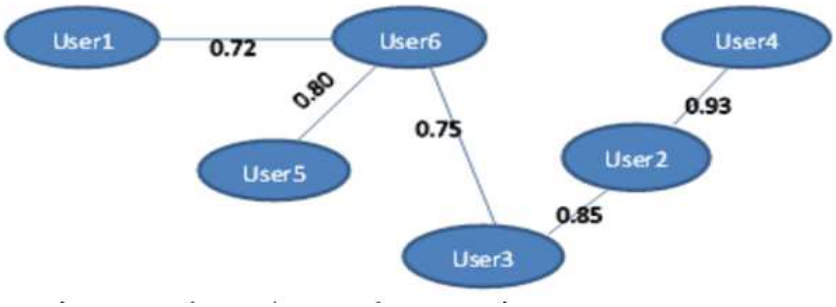

FIGURE 4.2 – Le graphe de similarité des utilisateurs

(a) ***Construction de graphe à partir des informations personnelles :*** Dans la vie quotidienne, nous remarquons que les personnes qui possèdent presque les mêmes âges et sexes ont des goûts similaires. Ainsi, nous avons pensé à créer un graphe qui les relie en utilisant les deux propriétés *(sexe, âge)*. La similarité Sse calcule comme suite :

$$similarite(u_p, u_k) = \frac{(|(sexe_p + sexe_k)\alpha + (age_p + age_k)\beta|)}{\alpha + \beta} \quad (4.1)$$

u_p , u_k representent les utilisateurs.

age_p, age_k representent l'age des utilisateurs u_p et u_k.

$sexe_p$, $sexe_k$ representent le sexe des utilisateurs u_p et u_k.

α , β sont respectivement les poids des propriétés sexe et age.

Avant le calcul de similarité, nous avons normalisé le sexe et l'âge. Un utilisateur est en relation avec un autre si la similarité est supérieure un seuil donné. β, α sont des poids pour l'âge et le sexe.

(b) ***Construire le graphe à partir des items communs (classement) :***

Dans cette étape, nous allons construire le graphe des utilisateurs en se basant sur leurs profils.

Pour un utilisateur U_i le système recherche les usagers U_j (j diffère de i) qui lui sont les plus similaires (par rapport aux items communs qui

se trouve dans leurs profils). Pour cela, nous avons utilisé la similarité entre deux usages avec la corrélation de Pearson. La formule suivante, nous donne cette valeur pour deux usagers A et B :

$$similarite(A,B) = \frac{\sum_{j \epsilon J}(v_{A,j} - \overline{v_A})(v_{B,j} - \overline{v_B})}{\sqrt{\sum_j (v_{A,j} - \overline{v_A})(v_{B,j} - \overline{v_B})}}. \quad (4.2)$$

A,B sont les utilisateurs.

$v_{A,j}$ represente l'evalution de l'utilisateur A à l'item j.

$\overline{v_A}$ represente le moyenne des evaluations de l'utilisateurs A.

J est l'ensemble des items evalué par les utilisateurs A est B.

(c) **Construire le graphe par hybridation entre les deux méthodes précédentes :**
S'il existe un lien direct entre l'utilisateur U_n et U_m dans les deux premiers graphes, par conséquence il y a une liaison dans ce dernier type de graphe. Autrement dit, la liaison n'aura pas lieu. Pour Construire le graphe par hybridation, nous avons utilisé l'algorithme02

Algorithme02 : Construction du graphe par hybridation

```
Require :
    Un, Um :Users
    // Gi est le graphe par informations personnelles
    // Gc est le graphe par classemnt
    Gi,Gc :graphe
    // Usersarray est un tableau d'utilisateurs
    Usersarray :Arraylist
while(Usersarray !=null)
{
   if( existe arc(Un,Um,Gi)&& existe arc(Un,Um,Gc))
        // Gh est le graphe par hybridation
        new graphe(un,um,Gh)
}
```

2. ***Les Items***

Afin de calculer la similarité entre les films, nous nous adaptons à un contexte fondé sur LOD, l'un des modèles les plus populaires en information classique : Le modèle vectoriel (VSM : Vector Space Model). Il existe

d'autres techniques se basant sur des modèles probabilistes ou booléens tel que Support Vector Machine, mais ils n'ont pas l'efficacité du VSM. Le modèle vectoriel de Salton en 1975 permet de calculer un degré (ou score) de ressemblance ou similarité entre la requête et le document. Les requêtes et les documents sont représentés par les vecteurs des poids des termes qui les constituent. Le score de ressemblance est exprimé comme la similarité entre le vecteur de la requête et le vecteur du document, généralement définie comme le cosinus entre les deux vecteurs.

Dans notre approche, nous nous intéressons aux VSM classique, habituellement utilisé pour la récupération de texte, pour traiter des graphes RDF. En un mot, nous représentons le graphe entier de RDF comme une matrice à trois dimensions où chaque tranche se rapporte à une propriété d'ontologie (écrivains, genre, acteur) et représente sa matrice de contiguïté(figure4.3).

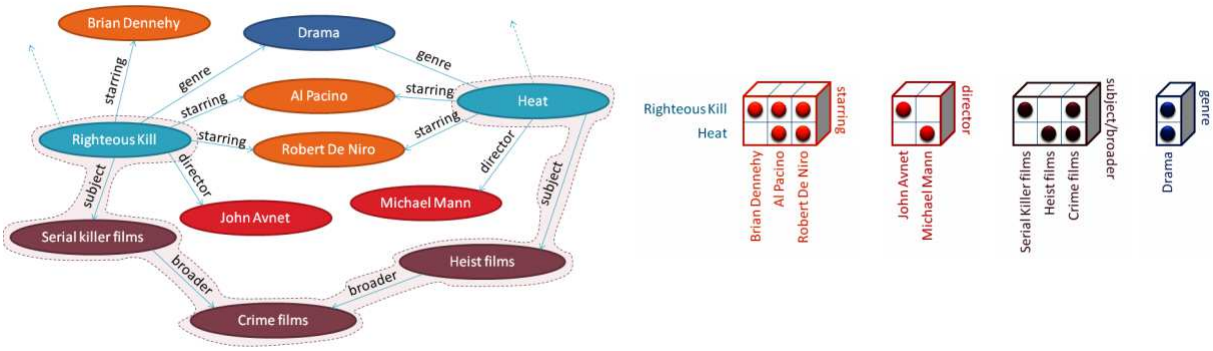

FIGURE 4.3 – La représentation matricielle d'un graphe RDF

Dans notre cas, chaque film est vu comme un vecteur, dont les composants référés à la $TFIDF$ (Term Frequecy-Inverse Document Frequency).

$$Le\ score\ de\ TFIDF = TF \times IDF \tag{4.3}$$

TF : terme frequency Comme son nom l'indique, cet indicateur permet de calculer la fréquence d'un mot ou expression dans un document. La formule est simple :

TF =Nbre de fois que le mot apparaît dans un document / nbre de mots dans le document. (4.4)

IDF : inverse terme frequency il introduit une notion de qualité du mot. C'est une mesure de l'importance du terme dans l'ensemble du corpus. Dans le schéma *TF-IDF*, vise à donner un poids plus important aux termes les moins fréquents.

$$idf = log\ (nbre\ de\ documents/nbre\ de\ documents\ contenant\ le\ terme) \tag{4.5}$$

3. ***Génération de modèle de données***
Les données recueillies à partir de la phase d'enrichissement seront également utilisées pour la génération du modèle de donnée (voir la Figure 4.4) qui est un graphe sémantique. En effet, des ensembles de données de LOD peuvent être vus en tant que graphes sémantiques où la connaissance est rattachée à une entité. Des ensembles de données sémantiques peuvent être employés comme entrée pour les systèmes de recommandations basés sur le contenu. En effet, un noeud donné représente un item.

Exemple :

Pour le film : *jumanji*, Nous pouvons utiliser les connaissances associées à l'entité correspondante (Genre de film : aventure) pour découvrir des items similaires disponibles dans le graphe. également, le modèle de données derrière des problèmes de filtrage collaboratif peut être aussi bien vu comme un graphe où les utilisateurs et les items sont les nœuds et les rétroactions d'utilisateurs sont des liens. La nature à base de graphe suggère des façons intéressantes pour modéliser un moteur de recommandation hybride.

Notre modèle de donnée est représenté par deux sous-graphes. Le premier sous-graphe représente le feedback contextuel des utilisateurs. Le deuxième sous-graphe représente la description sémantique de l'item.
Dans ce modèle, les films (nœuds *bleus*) sont des objets pour être recommandés aux utilisateurs (nœuds *rouges*). Les nœuds *verts* représentent les entités, venant de DBpedia.

(a) **Feedback contextuel de l'utilisateur**
Le contexte joue un rôle important en ce qui concerne la perception de l'utilité d'un objet pour un utilisateur. Cela peut influencer grandement la précision de la recommandation. Dans ce travail, trois dimensions de contexte sont exploitées : spatial, social et temporel.

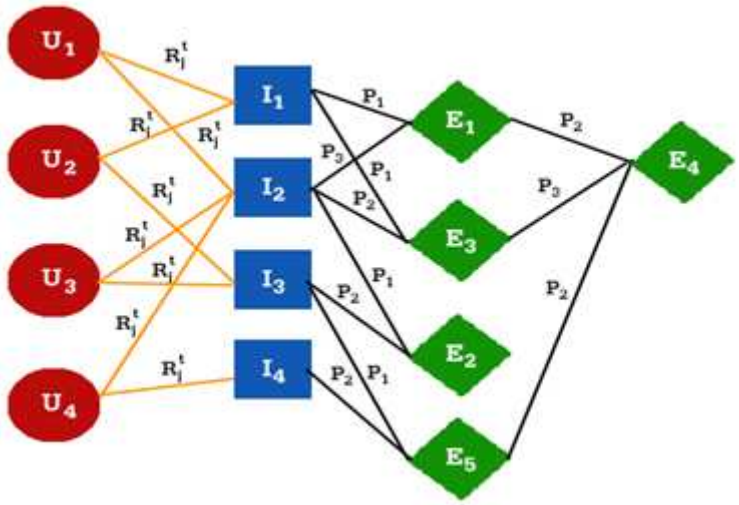

FIGURE 4.4 – Le modèle de donnée (graphe sémantique)

Définition 1 :
Formellement, le graphe du feedback contextuel de l'utilisateur est représenté par $G_c(U, I, R_j^t)$ où U est l'ensemble des utilisateurs, I est l'ensemble des items, et représente le valeur du feedback , t représente le contexte du feedback (time, location) et j la valeur du contexte (jour ou nuit pour le contexte time) .

(b) **Description sémantique de l'item**
La description sémantique des items se base sur l'exploitation des LOD (dbpedia).

Définition 2 Formellement, la description sémantique des items est modélisée par $G_i(I, P, E)$, où I et E représentent respectivement l'ensemble des items et leurs propriétés sémantiques, et P représente le poids de la propriété sémantique La génération du modèle de données se base sur l'algorithme 03.

4.4.2 Recommandation des items

Cette étape passe par deux phases. La génération des chemins et le filtrage des chemins.

Algorithme03 : La génération du modèle de données

```
Require :
    User U, Item I, E entity,C context,
    Number of user NU. Feedback F.
for i=1 to NU do
    new ARC(U ) ;
    new ARC(I ) ;
    new Nœud(U,I,C,F) ;
end for
```

4.4.2.1 Génération des chemins

Dans le but de faire une recommandation à un utilisateur, nous avons sélectionné les items non évalués par un utilisateur et extrait tous les chemins qui conduisent l'utilisateur vers ces items.

Exemple :

Entre l'utilisateur U_3 et l'item I_1, nous avons identifié huit chemins.
$(U_3, i_3, e_2, i_2, u_1, i_1)$,
(U_3, i_2, u_1, i_1),
$(U_3, i_2, e_3, e_1, i_1)$,
$(U_3, i_3, e_2, e_4, e_1, i_1)$,
$(U_3, i_3, e_5, e_4, e_1, i_1)$,
$(U_3, i_2, U_4, i_4, i_3, u_2, i_1)$,
$(U_3; i_3, i_4, e_5, e_4, e_1, i_1)$,
$(U_3, i_3, e_2, i_2, e_3, e_1, i_2, u_1, i_1)$.

4.4.2.2 Filtrage des chemins

Après avoir sélectionné tous les chemins conduisant d'un utilisateur vers un item non évalué, nous passons à la phase de filtrage des chemins. Nous prenons en considération les chemins qui ont une moyenne supérieure ou égale à un seuil donnée.

Il est à noter qu'il existe trois types de chemins.

1. ***Chemin collaboratif***
   C'est un chemin contenant uniquement des utilisateurs et des items (Figure 4.5). Pour identifier un chemin collaboratif, nous consultons le graphe du feedback contextuel de l'utilisateur (Gc) et nous vérifions l'existence d'une

relation entre les utilisateurs qui construisent le chemin collaboratif.

FIGURE 4.5 – Exemple de Chemin collaboratif

2. ***Chemin à base de contenu***
   Un chemin à base de contenu, est un chemin où l'utilisateur (concerné par la recommandation) et les items sont connectés par des entités (Figure 4.6). Pour identifier un chemin à base de contenu, nous utilisons la matrice des items (VSM) pour calculer la similarité entre les items qui construisent le chemin à base de contenu.

FIGURE 4.6 – Exemple de chemin à base de contenu

3. ***Chemin hybride***
   Un chemin hybride, est un chemin qui combine des items, des utilisateurs et des entités (Figure 4.7). Pour identifier un chemin hybride, nous consultons : i) le graphe du feedback contextuel de l'utilisateur (Gc) et nous vérifions l'existence d'une relation entre les utilisateurs qui construisent le chemin hybride, ii) la matrice des items (VSM) pour calculer la similarité entre les items qui construisent le chemin hybride.
   Exemple

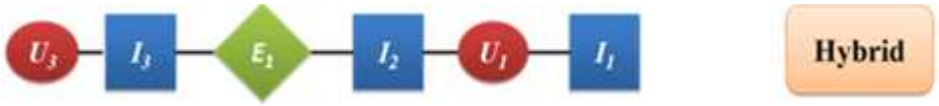

FIGURE 4.7 – Exemple de Chemin hybride

Pour les neuf chemins identifiés entre l'utilisateur U_3 et l'item I_1, et en considérant les poids de propriété sont $P_1 = 3$; $P_2 = 1$; $P_3 = 2$; $P_4 = 5$ et seul

égale à 2.5, nous avons retenu quatre chemins : C_1, C_6, C_7, C_8.

$C_1 : 5, 1, 3, 3, 2 = 14/5 = 2.8$ **;**
$C_2 : 1, 3, 2 = 6/3 = 2$;
$C_3 : 1, 1, 3 = 5/3 = 1.6$;
$C_4 : 1, 3, 2, 3 = 10/4 = 2,5$;
$C_5 : 5, 1, 2, 1, 3 = 12/5 = 2,4$;
$C_6 : 5, 3, 1, 1, 3 = 13/5 = 2,6$ **;**
$C_7 : 1, 5, 4, 5, 2, 4 = 21/6 = 3,5$ **;**
$C_8 : 5, 5, 1, 1, 1, 3 = 16/6 = 2,67$ **;**
$C_9 : 5, 1, 3, 1, 2, 1, 3, 2 = 18/8 = 2,25$;

4.5 Expérimentation

Nous avons effectué quelques expimentions afin d'évaluer la qualité des recommandations fournies. Dans cette section, nous conduisons la configuration expérimentale et nous fournissons une analyse complète des résultats expérimentaux.

4.5.1 Dataset

Nous avons mené plusieurs expériences sur l'ensemble de données populaires Movie Lens. L'ensemble de données, tiré du monde réel dans le domaine des films, contient 1 000 209 évaluations pour 3 883 films fournis par 6 040 utilisateurs.

Cette base de données inclus principalement :

- **Readme.txt (6 KB)** : contient la description de la base.
- **Movies.dat (168KB)** : informations sur 3 883 films. Format : *MovieID : :Title : :Genres*
- **Users.dat (132KB)** : informations sur 6 040 utilisateurs.Format : *UserID : : Gender : :Age : : Occupation : :Zip-code*
- **Rating.dat (24018 kB)** : 1 000 209 évaluations pour 3 883 films fournis par 6 040 utilisateurs. Format : *UserID : :MovieID : :Rating : :Timestamp*

Les données de MovieLens visent principalement à évaluer les systèmes de recommandation collaboratifs dans le domaine du film. étant donné que notre approche repose sur une recommandation basée sur le contenu et afin d'utiliser ces ensembles de données pour tester les performances de nos algorithmes, nous avons relié les ressources représentées par MovieLens à celles de DBpedia.

4.5.2 Environnent d'expérimentation

Pour évaluer les performances de notre système, nous avons développé un prototype (figure 4.8) à base de java (jdk 1.8.0) et sous l'environnement Netbeans 7.0, 8. Nous avons conduit un ensemble d'expérimentation pour évaluer la performance et la fiabilité de ce système. Ces expérimentations sont menées sur une plateforme, ayant un processeur Core i7 avec 08 GO de RAM et sous le système d'exploitation Windows 8

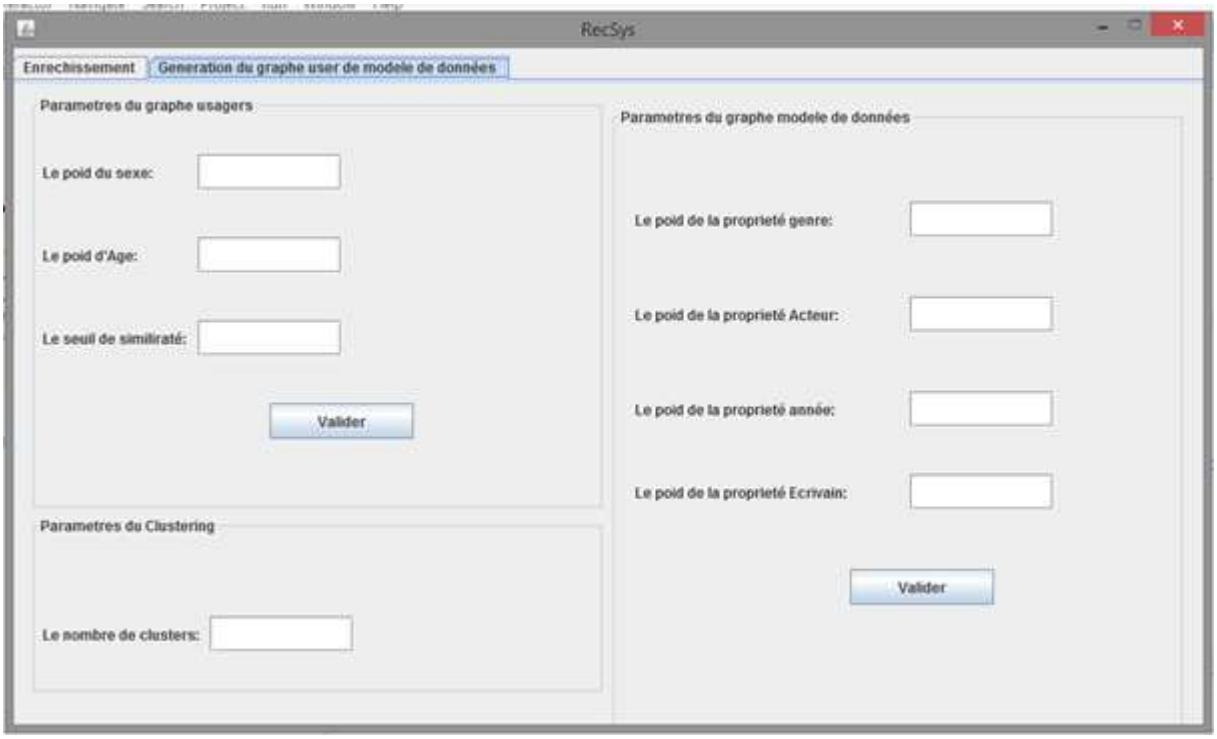

FIGURE 4.8 – Interface principale de notre système de recommandation

4.5.3 L'impact des méthodes de filtrage collaboratif sur le système de recommandation

Cette expérimentation est faite afin de comparer entre les résultats de l'évaluation de notre système tout en jouant avec les paramètres de du graphe utilisateur(figure4.9). 3 tests ont été éffectué :

- Dans le premier test, nous avons utilisé la methode qui repose sur les informations personnelles (age et sexe) avec poids=1 et un seuil de similaritéde 0,5

- Lors du deuxieme test, nous avons utilisé lamethode de classement avec un seuil de 0,5
- Enfin, on considere une hybridation des deux précedentes methodes avec un seuil de 0,5 préalablement (0.5 pour chaque méthode).

D'aprés ce test nous remarquons que la combinaison entre les informations personnelles et la methode de classement donne les meilleurs résultats

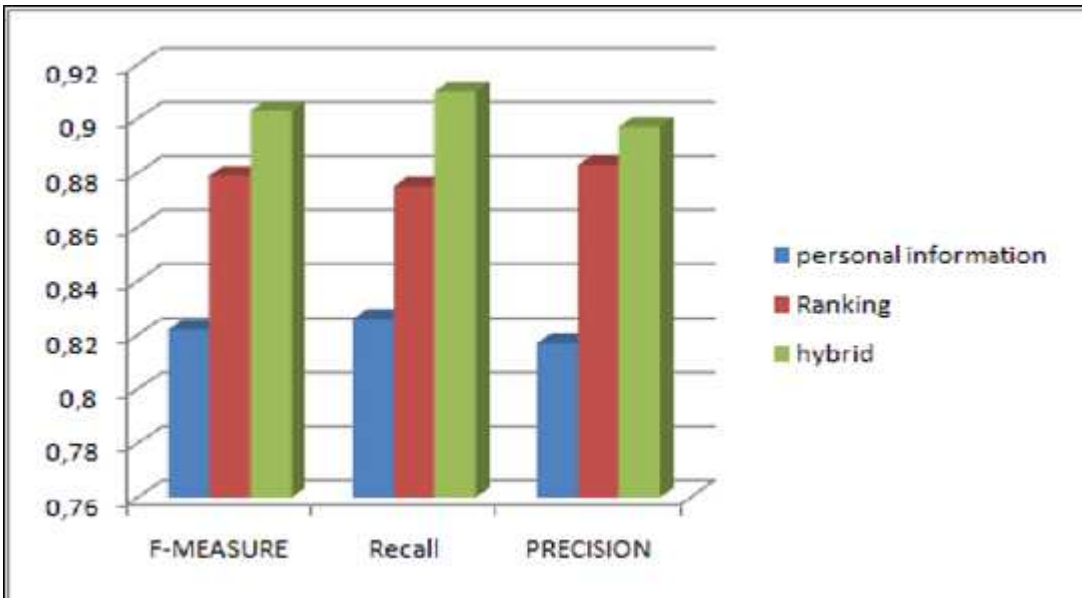

FIGURE 4.9 – L'évaluation des méthodes de filtrage collaboratif

Un 4ème test à été effectué pour identifier les meilleurs combinaison(figure4.10)

4.5.4 L'impact des méthodes de filtrage à base de contenu sur le système de recommandation

Nous avons comparé entre les résultats de l'évaluation de notre système tout en modifiant les paramètres de l'algorithme VSM(figure4.11) .

- Dans la première expérimentation les poids des propriétés écrivain, acteur et genre sont égaux à 1.
- Dans la deuxième expérimentation le poids de la propriété écrivain est égale à 1, 2 pour la propriété acteur et 5 pour la propriété genre.

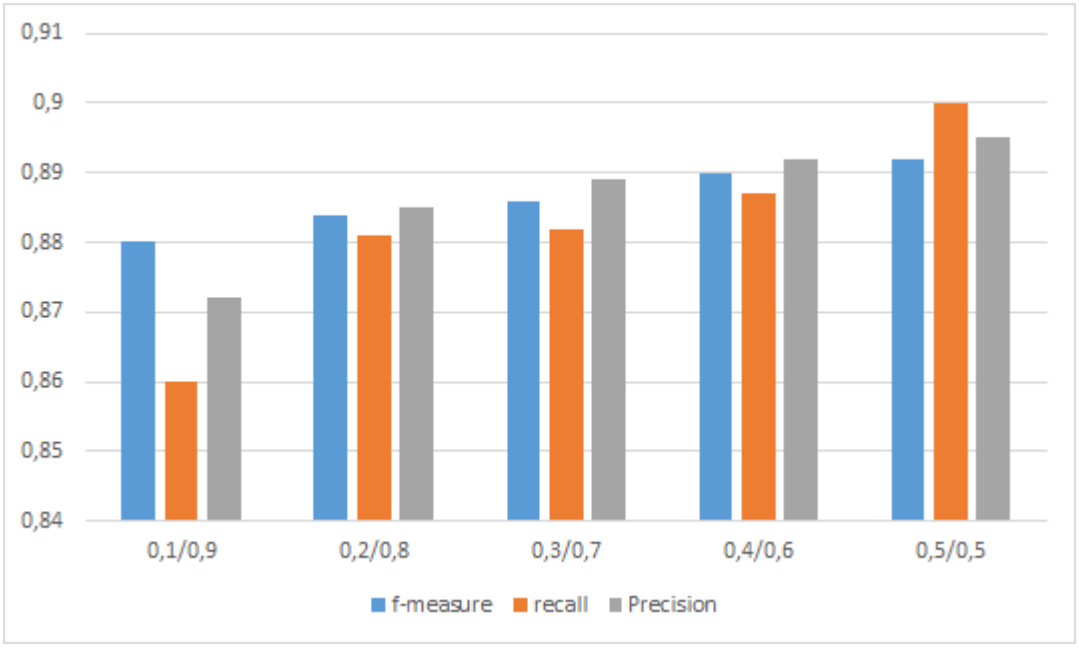

FIGURE 4.10 – L'évaluation de l'hybridation du graphe d'informations personnelles et le graphe de classement

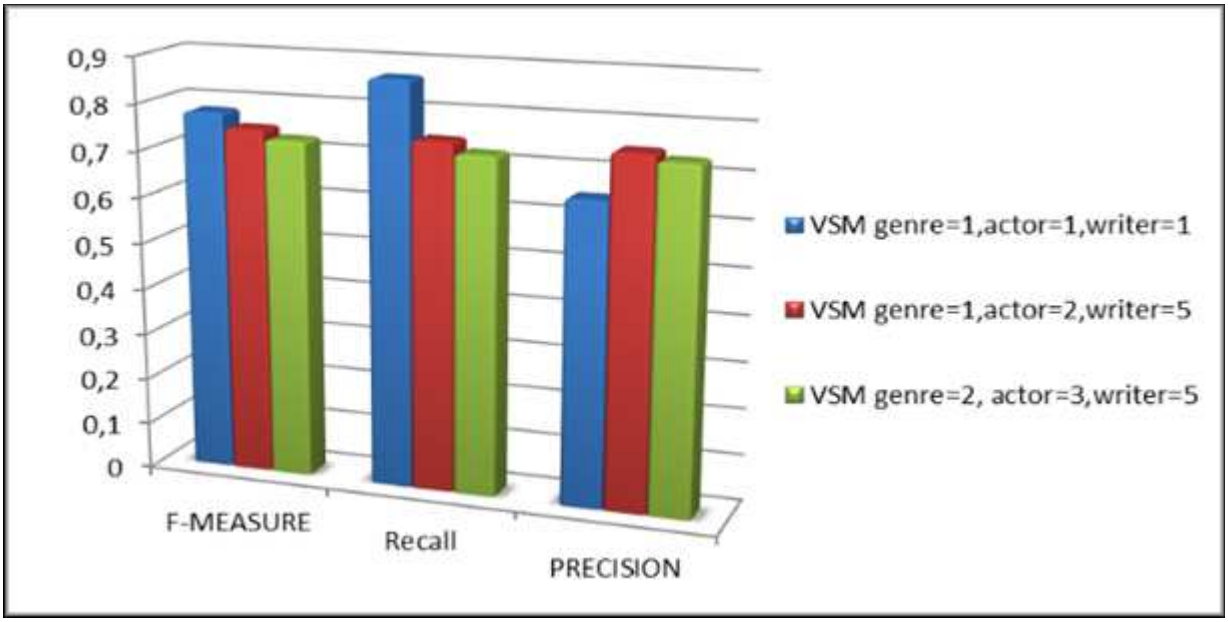

FIGURE 4.11 – L'évaluation du VSM

- Dans la deuxième expérimentation le poids de la propriété écrivain est égale à 2, 3 pour la propriété acteur et 5 pour la propriété genre.

Les résultats de cette expérimentation sont présentés sur la Figure 4.11. Nous constatons que le changement des paramètres de VSM influe sur la recommandation

4.5.5 Impact du nombre de cluster sur la recommandation

Dans la Figure 4.12, nous notons que la modification du nombre de grappes d'utilisateurs (k = 1 à 5) affecte la recommandation. Sur la base de cette observation, nous avons choisi k = 2 pour le reste des expériences.

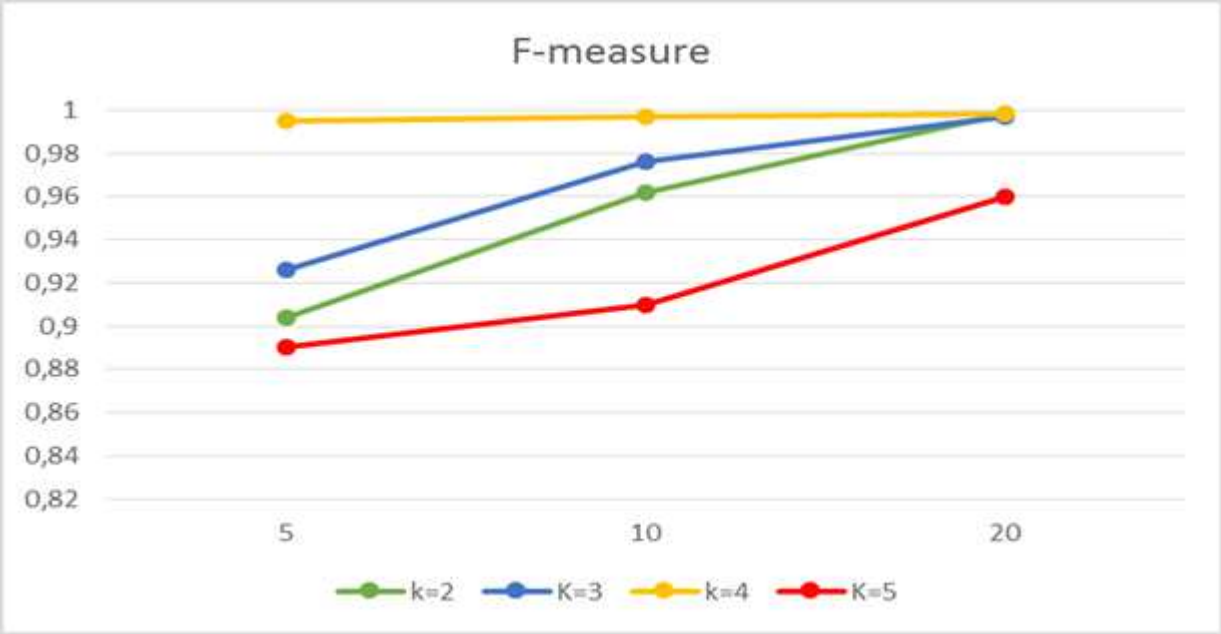

FIGURE 4.12 – Impact du nombre de cluster sur la recommandation

4.5.6 Comparaison entre les différentes techniques de regroupement

La Figure 4.13 compare entre les différentes techniques utilisées pour regrouper les utilisateurs et les éléments respectivement : (Kmeans, Kmeans), (Kmeans, VSM), (Graphique, Kmeans), (Graphique, VSM). Il est évident que les meilleurs résultats sont obtenus lors de l'utilisation de Kmeans pour le regroupement d'utilisateurs et de VSM pour le regroupement d'éléments.

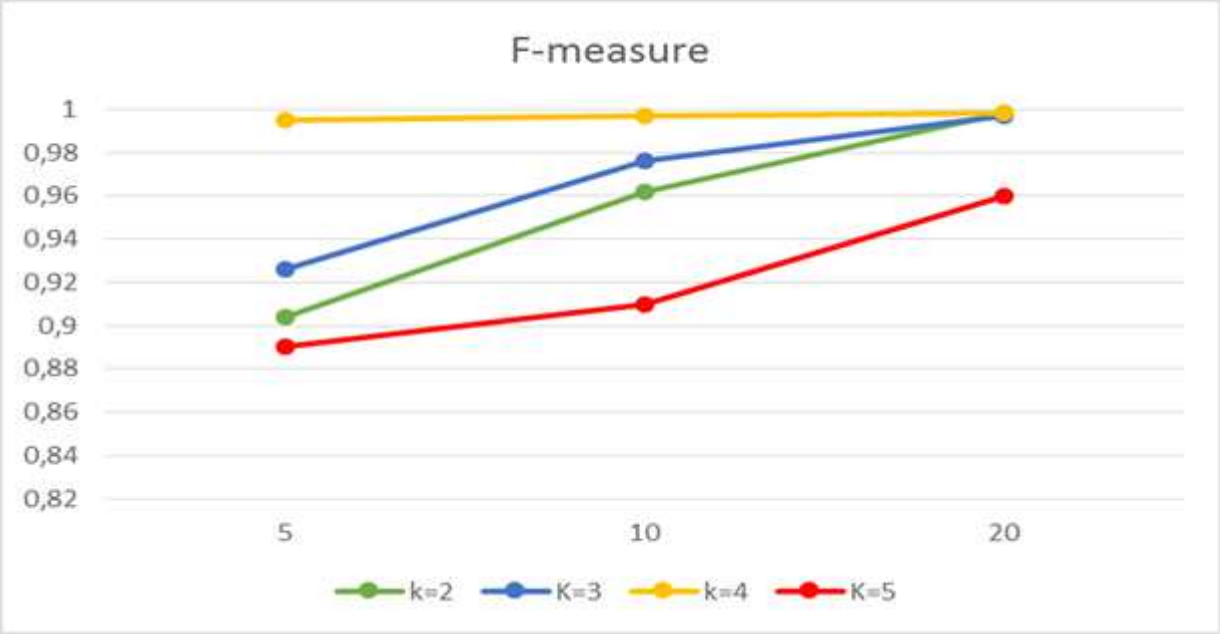

FIGURE 4.13 – Comparaison entre les différentes techniques de regroupement

4.6 Conclusion

L'utilisation des technologies sémantiques pour améliorer les systèmes de recommandations est un domaine de recherche naissant. Dans ce chapitre, nous avons proposé une nouvelle méthode hybride qui traite le feedback explicite via un graphe de données qui est constitué de tous les utilisateurs et les items du système. Nous avons montré que les données liées peuvent être utilisées pour « combler les lacunes » des algorithmes de filtrages. Cela nous permet d'acquérir les données nécessaires pour atténuer les problèmes afin de formuler des recommandations pertinentes aux nouveaux utilisateurs et de nouveaux items, ainsi que l'augmentation de la valeur des recommandations d'une façon générale.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

L'expansion de l'Internet et du nombre d'applications basées sur le Web, est associée à une prolifération d'information ou d'items dont le volume ne cesse de croître. Devant cette profusion et cette surcharge d'items, l'utilisateur peine à repérer l'information pertinente qui correspond le plus à ses besoins. Dans ce contexte, les systèmes de recommandation ont été développés en vue de faciliter l'accès à ces items pertinents. Leur objectif est d'anticiper les besoins de l'utilisateur en lui fournissant des recommandations d'items jugés pertinents par rapport à ses goûts.

La dernière décennie a été marquée par un large déploiement des systèmes de recommandation dans différents champs d'application. Il existe une variété de techniques de filtrage dans les systèmes de recommandation, tel que le filtrage basé sur le contenu, le filtrage collaboratif et d'autre qui ont été relatés dans le chapitre 2 du présent document. Toutefois, des problèmes subsistent toujours, parmi lesquels on peut citer le démarrage à froid et la rareté des votes, manque de données sur l'item ou l'utilisateur. Afin de trouver quelques réponses à ces problèmes, nous avons exploité le domaine de web sémantique plus précisément, les données liées qui ont été clairement élucidées dans le chapitre 3 de notre travail. L'utilisation des technologies sémantiques pour améliorer les systèmes de recommandations est un domaine de recherche naissant, nous nous sommes intéressés a étudié et analyser des approches qui visent à utiliser les données liées dans les systèmes de recommandation dans l'état de l'art (chapitre 4).

Nous nous sommes inspirés des travaux de Osttuni ([Ostuni et al, 2012], [Otsuni et al, 2013]) qui utilisent les données liées pour enrichir les items à travers VSM dans le cadre d'un filtrage à base de contenu, tout en offrant les meilleures N

recommandations par rapport aux feedbacks implicites des utilisateurs.

Dans cette thèse, nous proposons une nouvelle approche hybride qui combine le filtrage collaboratif avec le filtrage à base de contenu tout en collectant des données auprès des LOD afin de minimiser les problèmes mentionnés préalablement.

L'approche traite les évaluations via un graphe des données qui se compose de tous les utilisateurs et les éléments du système. Pour le filtrage à base de contenu, nous avons créé un VSM pour les items. En ce qui concerne le filtrage collaboratif, nous avons construit un graphe entre les différents utilisateurs en se basant sur le calcul de similarité entre eux.

5.2 Perspectives

Ce travail ouvre la voie à de nouvelles perspectives intéressantes qui restent à explorer, et qui vont contribuer à l'évolution des propositions que nous avons réalisées dans le cadre de cette thèse.

1. D'abord, nous envisageons d'étendre notre approche par l'utilisation d'autres sources de données liées tel que DBTrope, Freebase et LinkedMDB, dans le but d'enrichir et d'améliorer la description du profile utilisateur et celle des items.

2. Ensuite, nous nous intéresserons à la combinaison entre le feedback implicite et explicite des utilisateurs pour améliorer la précision de la recommandation

3. Enfin, nous planifions d'évaluer les performances de notre approche avec d'autres Dataset

Bibliographie

[Aciar et al , 2007]Aciar, S., Zhang, D., Simoff, S., &Debenham, J. InformedRecommender : BasingRecommendations on Consumer Product Reviews. IEEE Intelligent Systems 22, 39-47, 2007.

[Adomavicius et al, 2005] Adomavicius, G. and Tuzhilin, A. Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6) :734–749, 2005.

[Ahn et al, 2007]Ahn, J., Brusilovsky, P., Grady, J., He, D., & Syn, S.. Open User Profiles for Adaptive News Systems : Help or Harm ? 16th International Conference on World Wide Web (pp. 11-20). ACM, 2007.

[Anderson , 2011] Chris Anderson. Recommender systems for eshops. Business Mathematics and Informatics paper.Amsterdam :Faculty of Sciences Vrije Universiteit, 2011, 34 p.

[Benhmidi , 2011] H.BENHMIDI .Chapitre I - Les ontologies. p 13, 2011 [Berners-Lee , 2001]Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, Mai 2001.

[Breese et al ,1998] Breese J.S., Heckerman D., Kadie C., "Empirical analysis of predictivealgorithms for collaborative filtering", Proceedings of the 14th Annual-Conference on Uncertainty in Artificial Intelligence, p. 43-52, 1998.

[Burke, 2002] Robin BURKE, Hybrid Recommender Systems : Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4) :331–370, 2002.

[Chen et al, 1998] Chen, L., &Sycara, K. WebMate : A Personal Agent for Browsing and Searching. 2nd International Conference on Autonomous Agents (pp. 9-13). New York : ACM Press, 1998.

[Das et al., 1997] Das, D., Hill, W., Stead, L. &Wittenburg, K. (1997) Group Asynchronous Browsing on the World Wide Web. In proceedings of the 6th International Conference on the World Wide Web (WWW'6), 1997

[Degemmi et al , 2007]Degemmis, M., Lops, P., &Semeraro, G. A Content-collaborative Recommender that ExploitsWordNetbased User Profiles for Neighborhood Formation. User Modeling and User- Adapted Interaction : The Journal of Personalization Research (UMUAI), 217-255 , 2007.

[Eirinaki et al, 2003] Eirinaki, M., Vazirgiannis, M., &Varlamis, I. SEWeP : Using Site Semantics and a Taxonomy to enhance the Web Personalization Process. Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 99-108), 2003.

[Furnas et al., 1995] Furnas, G., Hill, W., Rosenstein, M. & Stead, L. (1995) Recommending and evaluating choices in a virtual community of use. In Proceedings of ACM CHI'95, pages 194–201, 1995.

[Goldberg et al., 1992] Goldberg, D., Nichols, D., Oki, M.B., & Terry, D. Using collaborative filtering to weave an information tapestry. Commun. ACM, 35(12), 61-70, 1992.

[Heath and Bizer , 2011] Tom Heath & Christian Bizer .Linked Data : Evolving the Web into a Global Data Space.[en ligne].(Publié 2011)

[Heitmann et al, 2010] Benjamin Heitmann & Conor Hayes. Using Linke Data to Build Open, Collaborative Recommender Systems. Digital Enterprise Research Institute National University & Galway Galway, Ireland, 2010,P 6 (76-81).

[Jannach et al, 2010] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. Recommender Systems :An Introduction. Cambridge University Press, New York, NY, USA,1st edition, 2010.

[Konstanet al., 1997] Konstan, J.A., Miller, B. &Riedl, J.Experiences with GroupLens : Making Usenet useful again, Proceedings of the 1997 Usenix Winter Technical Conference,1997.

[Ko et al,2014] Han-Gyu Ko & Eunae Kim & In-Young Ko & Deokmoon Chang . Semantically-based Recommendation by using Semantic Clusters of Users' Vie-

wing History. Department of Computer Science & Division of Web Science and Technology & Korea Advanced Institute of Science and Technology & Advanced Institute of Technology, Daejeon & Seoul, Republic of Korea, 2014, P 5 (83-87)

[Laublet , 2003] P. Laublet C. Reynaud J. Charlet .Sur quelques aspects du Web sémantique . Université de Paris-Sorbonne, Université Paris-Sud, Université Paris-X Nanterre, 2003, 20 p

[Liberman, 1995] Liberman, H ,Letizia : An Agent that Assists Web Browsing. International Joint Conference on ArtificialIntelligence (IJCAI-95) (pp. 924-929). Montreal, Canada : Morgan Kaufmann publishers Inc,1995 .

[Maes et al, 1995]Maes, P. &Shardanand, U, Social information filtering : algorithms for automating "word of mouth", Proceedings of the SIGCHI conference on Human factors in computing systems. Denver, Colorado, United States : ACM Press/Addison-Wesley Publishing Co, 1995.

[Malone et al.,1987] Malone, T.W., Brobst, S.A., Cohen, S.A., Grant, K.R. &Turbak, F.A. (1987) Intelligent information des systèmes de partage. Communications of the ACM, 30 (5) :390-402, mai 1987.

|Magnini et al, 2001] Magnini, B., &Strapparava, C. Improving User Modelling with Content-based Techniques. 8thInternational Conference of User Modeling, (pp. 74-83).

[Margaritiset al., 2003] Margaritis, K.G. &Vozalis, E. Analysis of Recommender Systems' Algorithms. 6th Hellenic European Conference on Computer Mathematics & its Applications (HERCMA), 2003

[Mestiri , 2007] Mohamed Amine Mestiri .Chapitre 2 : état de l'art.[en ligne].(2007) Disponible sur : <http ://theses.ulaval.ca/archimede/ fichiers/24629/ch02.html />.

[Mladenic , 1999] Mladenic, D. Machine learning used by PersonalWebWatcher. ACAI-99 Workshop on Machine Learning and Intelligent Agents, 1999

[Middleton et al , 2004] Middleton, S., Shadbolt, N., & De Roure, D. Ontological User Profiling in Recommender Systems. ACM Transactions on Information Systems, 54-88, 2004.

[Naak, 2009]Naak, A. Papyres : Un système de gestion et de recommandation d'articles de recherche. Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maîtrise ès Sciences en Informatique, Montréal, Canada, 2009

[Ostuni et al, 2012] Tommaso Di Noia & Roberto Mirizzi & Vito Claudio Ostuni & Davide Romito &Markus Zanker . Linked Open Data to support Content-based Recommender Systems.Italy :Politecnico di Bari & Austria :University Klagenfurt, 2012,P 8.

[Ostuni et al, 2013] Vito Claudio Ostuni & Tommaso Di Noia & Eugenio Di Sciascio & Roberto Mirizzi. Top-N Recommendations from Implicit Feedback leveraging Linked Open Data.Italy : Polytechnic University of Bari, 2013, P8 (85-92).

[Peska et al,2013] Ladislav Peska & Peter Vojtas. Enhancing Recommender System with Linked Open Data. Faculty of Mathematics and Physics & Charles University in Prague & Malostranske Namesti 25, Prague, Czech Republic, 2013, P 12 (483-494).

[Piamrat et al, 2009] K. Piamrat, C. Viho, J.-M. Bonnin, and A. Ksentini. Quality of Experience Measurements for Video Streaming over Wireless Networks. In Sixth International Conference on Information Technology : New Generations, 2009.ITNG '09, pages 1184 –1189, April 2009.

[PLU , 2011] Julien Plu .Introduction au Web sémantique.[en ligne].(Publié le 21 avril 2011) Disponible sur :<http ://jplu.developpez.com/tutoriels/web-semantique/introduction>.

[Polytec , 2012] Polytec Lyon .Web sémantique sur les systèmes embarqués.[en ligne].(Dernière date de mise à jour : 15 mai 2012) Disponible sur :<http ://pagesperso.lina.univ-nantes.fr/ prie-y/archives/VEILLE- 2009 2012/2012/wsembarq/2.html/ >.

[Ragone et al, 2017] Azzurra Ragone & Paolo Tomeo & Corrado Magarelli & Tommaso Di Noia & Matteo Palmonari & Andrea Maurino & Eugenio Di Sciascio , Schema summarization in Linked-Data-based feature selection for recommender systems , SAC'17, April 3-7, 2017, Marrakesh, Morocco, 2017

[Rao et al , 2008] Rao, N.K., and Talwar, V.G. Application domain and functional classification of recommender systems a survey. In Desidoc journal of library and information technology, vol 28, n3, 17-36, 2008.

[Resnicket al., 1997] Resnick, P. & Varian, H.R,Recommender systems. Communications of the ACM, 40(3) :56–58, 1997.

[Rochlitz, 1992] ROCHLITZ : Dans le flou artistique. élements d'une théorie de la « rationalité esthétique ».Bouchindhomme et Rochlitz, pages 203–238, 1992

[Su et al, 2009] Su, X., and Khoshgoftaar, T.M. A survey of collaborative filtering techniques. In Adv. in Artif. Intell. 2009

[Terveenet al., 1997]Terveen, L.G., Creter, J., Hill, W.C., McDonald, D. &Amento, B. Building Task-Specific Interfaces to High Volume Conversational Data, CHI'97, 1997.

[Yang et al, 2013] Rui Yang & Wei Hu & and Yuzhong Qu. Using Semantic Technology to Improve Recommender Systems Based on Slope One. Department of Computer Science and Technology, Nanjing University, China, 2013, P 13 (11-23)

[Yates et al., 1999] Baeza-Yates, R., & Ribeiro-Neto, B.. Modern Information Retrieval. Addison-Wesley, 1999.