Système de recommandation sémantique enrichi : application au domaine du e-marketing

Baba Mbaye

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
PRÉPARÉE AU LABORATOIRE ELLIADD - EA 4661

École doctorale 592
Lettres, Communication, Langues, Art

Doctorat en Sciences de l'Information et de la Communication

Par

Baba MBAYE

SYSTÈME DE RECOMMANDATION SÉMANTIQUE ENRICHI. APPLICATION AU DOMAINE DU E-MARKETING

Thèse présentée et soutenue à Montbéliard le 25 avril 2022

Composition du Jury :

Mme. BALICCO, Laurence	Professeure à l'Université Grenoble Alpes	Examinatrice
Mme. CALABRETTO, Sylvie	Professeure à l'INSA de Lyon	Rapportrice
M. CARO, Stéphane	Professeur à l'Université Bordeaux Montaigne	Président du jury
M. LAZAR-FAVORY, Loïc	Président de l'entreprise Effet B Lyon	Invité
M. ROXIN, Ioan	Professeur à l'Université de Franche-Comté	Directeur de thèse
M. SALEH, Imad	Professeur à l'Université Paris 8	Rapporteur
M. TAJARIOL, Federico	MCF HDR à l'Université de Franche-Comté	Co-encadrant

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
PRÉPARÉE AU LABORATOIRE ELLIADD - EA 4661

École doctorale 592
Lettres, Communication, Langues, Art

Doctorat en Sciences de l'Information et de la Communication

Par

Baba MBAYE

SYSTÈME DE RECOMMANDATION SÉMANTIQUE ENRICHI. APPLICATION AU DOMAINE DU E-MARKETING

Thèse présentée et soutenue à Montbéliard le 25 avril 2022

Composition du Jury :

Mme. BALICCO, Laurence	Professeure à l'Université Grenoble Alpes	Examinatrice
Mme. CALABRETTO, Sylvie	Professeure à l'INSA de Lyon	Rapportrice
M. CARO, Stéphane	Professeur à l'Université Bordeaux Montaigne	Président du jury
M. LAZAR-FAVORY, Loïc	Président de l'entreprise Effet B Lyon	Invité
M. ROXIN, Ioan	Professeur à l'Université de Franche-Comté	Directeur de thèse
M. SALEH, Imad	Professeur à l'Université Paris 8	Rapporteur
M. TAJARIOL, Federico	MCF HDR à l'Université de Franche-Comté	Co-encadrant

ECOLE DOCTORALE
LECLA

Titre : Système de recommandation sémantique enrichi. Application au domaine du e-marketing.

Mots clés : système de recommandation, interprétation d'information, extraction de connaissances, outil d'annotation sémantique, apprentissage automatique

Résumé : Cette thèse est ancrée dans les champs des sciences de l'information et de la communication, de l'ingénierie des connaissances et appliquée au domaine du e-marketing et plus précisément aux enquêtes mystères. Une enquête mystère est une visite en point de vente effectuée par un enquêteur. La visite en point de vente a pour but de mesurer la qualité de l'accueil, la qualité du conseil et le respect des consignes de vente.

Les rapports produits par les « enquêtes mystères » et les études de satisfaction sont des notes, des classements et des verbatims issus de questions ouvertes et fermées. Les commanditaires des enquêtes utilisent ces rapports pour définir des plans d'actions pour améliorer la qualité et l'efficacité du service, allant des changements organisationnels jusqu'à la mise en place de formations spécifiques pour augmenter les compétences du personnel.

Ainsi, l'enjeu de cette thèse est d'exploiter les technologies du web sémantique et de l'apprentissage automatique pour mettre en place un système de recommandation capable d'analyser et de partiellement interpréter les données collectées.

Pour ce faire, nous nous intéressons à l'élaboration et l'expérimentation de nouvelles approches d'annotation sémantique de données. La mise en place de ce système d'interprétation nécessitera une base de connaissances qui sera alimentée par ces annotations (métadonnées). Nous utilisons aussi les technologies de l'apprentissage automatique pour enrichir le traitement de la recommandation, pour améliorer la pertinence de la prédiction et pour classifier automatiquement les items à recommander pour mieux guider l'expert dans sa prise décision finale sur les axes à améliorer au sein des points de vente.

UBFC
UNIVERSITE
BOURGOGNE FRANCHE-COMTÉ

Université de Bourgogne
32, avenue de l'Observation
25200 Besançon

ECOLE DOCTORALE
LECLA

Title : Enriched semantic recommendation system. Application to the e-marketing domain.

Keywords : recommendation system, interpretation of information, knowledge extraction, semantic annotation tool, machine learning

Abstract : This thesis is anchored in the fields of Information and Communication Sciences, knowledge engineering and applied to the field of e-marketing and more precisely to mystery shopping. A mystery survey is a point-of-sale visit carried out by an investigator. The purpose of the point of sale visit is to measure the quality of the reception, the quality of the advice and the respect of the sales instructions.

The reports produced by « mystery surveys » and satisfaction studies are notes, rankings and verbatim reports based on open and closed questions. Survey sponsors use these reports to define action plans to improve service quality and efficiency, ranging from organizational changes to the implementation of specific training to increase staff skills. Thus, the challenge of this thesis is to exploit the technologies of the semantic web and automatic learning to set up a recommendation system capable of analyzing and partially interpreting the collected data.

To do this, we are interested in developing and testing new approaches to semantic annotation of data. The implementation of this interpretation system will require a knowledge base that will be fed by these annotations (metadata). We also use machine learning technologies to enrich the recommendation processing, to improve the relevance of the prediction and to automatically classify the items to be recommended in order to better guide the expert in his final decision making on the areas to be improved within the points of sale.

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

Université de Bourgogne
32, avenue de l'Observation
25200 Besançon

À mes très chers parents, pour l'amour qu'ils m'ont toujours donné, leurs encouragements et toute l'aide qu'ils m'ont apportée durant mes études.

« *La valeur d'une éducation universitaire n'est pas l'apprentissage de nombreux faits, mais l'entraînement de l'esprit à penser.* »

Albert Einstein

Remerciements

Tout d'abord, je souhaite exprimer ma reconnaissance et remercier mes directeurs de thèse, Pr. Ioan Roxin et M. Federico Tajariol qui m'ont accueillie dans leurs équipes de recherche et initier au métier de la recherche scientifique. Très patient et rigoureux dans leurs encadrements, je les remercie pour leurs confiances, leurs conseils et pour leur soutien afin de mener à bien mes travaux de recherche, toujours avec pédagogie et le souci du détail. Nos discussions tout au long de ces années m'ont aidé à nourrir ma réflexion en qualité de chercheur.

Je tiens à remercier les membres du jury, d'avoir accepté d'évaluer mes travaux de recherche et pour l'intérêt qu'ils ont porté à mon travail. J'adresse mes remerciements à Mme. Sylvie Calabretto et à M. Imad Saleh pour leur rôle de rapporteurs, à Mme. Laurence Balicco pour avoir accepté d'évaluer mon travail ainsi qu'à M. Stephane Caro d'avoir assuré la présidence du jury. Leurs questions, remarques et suggestions formulées dans les rapports et lors de la soutenance me sont importantes pour améliorer ma recherche.

Cette thèse a été menée dans le cadre d'une convention industrielle de formation par la recherche (CIFRE) en collaboration avec l'entreprise Effet B que je remercie beaucoup, notamment à son directeur général, Loïc Favory pour sa confiance et sa détermination pour la réussite de mes travaux de recherche. Mes remerciements vont aussi à mes collègues de l'entreprise Effet B pour leur amitié et leur soutien moral.

J'adresse mes remerciements aux collègues du pôle CCM du laboratoire ELLIADD que j'ai côtoyés pour les échanges que nous avons eus. Tous mes encouragements à Clément Auboeuf et He Li pour le reste du chemin à parcourir. Je remercie Jean-Marie Leygonie et Aymeric Bouchereau qui m'ont apporté leur aide pour la relecture du manuscrit.

Enfin, je remercie ma famille, notamment mes parents, mes frères et ma sœur pour leur soutien.

Je ne pourrais pas terminer sans remercier ma chère compagne, d'avoir toujours été à mes côtés et d'avoir partagé tous les moments, les bons comme les moins bons, que j'ai connus pendant cette période.

Introduction

Les plateformes numériques sont des sources importantes d'acquisition d'information fournit par ses usagers. À titre d'exemple, sur la plateforme Facebook [1], le nombre d'utilisateurs est passé de cent millions en 2008 à deux milliards en 2017, ce qui augmente considérablement la quantité d'information publiée sur la plateforme.

Cette augmentation de la quantité d'information est à l'origine de la manière d'utilisation de ces plateformes : les usagers ne se limitent plus à la consommation des informations publiées, ils sont aussi auteurs de ces contenus. Cette création de contenus va au-delà des publications faites. Les usagers ont la possibilité d'interagir entre eux, ainsi qu'avec l'ensemble des contenus qu'ils consultent. Ces interactions sont sur plusieurs formes, comme par leurs avis sur un ensemble de produits d'un commerce en ligne, par le renseignement d'un questionnaire numérique de satisfaction à la suite d'un achat d'un produit auprès d'un magasin ou encore par l'acquisition de résultats d'enquêtes à la suite d'une vague d'enquête mystères dans des points de vente. Dans cette thèse, nous allons nous intéresser à la collecte et au traitement des données issues des résultats d'enquêtes mystères.

Les enquêtes mystères constituent en marketing un moyen pour vérifier concrètement la bonne commercialisation des produits d'un ou plusieurs points de vente. Elles peuvent aussi constituer un outil d'orientation de thématiques de contrôles, utile pour la programmation d'une politique globale d'animation d'un réseau de vente. Ces enquêtes mystères permettent aux entreprises de déceler le plus en amont possible les dysfonctionnements éventuels de la chaîne de commercialisation des produits, de mettre en évidence les bonnes et mauvaises pratiques et de détecter les risques de ventes abusives ou inadaptées de produits. Elles permettent aussi aux entreprises, d'acquérir une mesure concrète de la qualité d'un service offert afin de l'améliorer, en modifiant leur offre ou en proposant des formations à leurs employés. La répétition des campagnes d'enquêtes mystères permet de réduire la portée de certains biais : sur la durée, des informations sur l'évolution des pratiques commerciales observées pourront être collectées utilement par le régulateur.

Les enquêtes mystères sont réalisées par de « faux » clients recrutés et missionnés par les entreprises pour évaluer leur réseau de ventes. Elles sont utilisées dans beaucoup de sec-

1. https ://kinsta.com/fr/blog/statistiques-facebook/

teurs économiques (e.g. Automobile, Banques). Les enquêtes mystères sont complémentaires aux enquêtes de satisfaction qui donnent une mesure d'opinion ainsi que des réclamations dénonçant des éventuels dysfonctionnements.
Des interactions régulières entre des clients mystères et le personnel d'un réseau de vente deviennent des rituels qui renforcent l'appartenance collective à une marque et favorisent la réalisation d'objectifs communs. L'ensemble du processus de vente au détail devient une chaîne d'interactions, une série d'actions et des réponses dans des situations sociales et des échanges économiques avec des attentes spécifiques pour chaque participant et des règles de comportement détaillées, bien que généralement non écrites. Le client mystère représente tous les clients et doit agir conformément à tous les aspects habituels. Le personnel doit répondre aux attentes rituelles prescrites pour que le client mystère soit satisfait du service. Le défaut de message d'accueil approprié, par exemple, peut créer un « rituel gâté » avec des conséquences et des coutumes prescrites.
L'objectif d'une enquête mystère est dans un premier temps de connaître ses clients et ses prospects, ce qui est très important pour déterminer par exemple les canaux de communications à adopter. Elle a aussi pour objectif d'identifier les produits et les services qui intéressent les consommateurs. En résumé ces enquêtes permettent de comprendre le comportement de la clientèle et des prospects dans le but de prendre des décisions visant à améliorer l'offre commercial.

Ces dernières années avec la croissance significative des plateformes informatiques, la plupart des entreprises de pilotage d'enquêtes mystères se sont orientées vers le numérique, dans le but de déléguer une partie du processus et du traitement de ces enquêtes mystères. Cette numérisation est un moyen important pour ces entreprises sur le fait de gagner du temps. Parmi les entreprises qui utilisent le numérique pour le pilotage des enquêtes mystères on peut citer : Converso [2], Orphée [3], Swiss audit Shpo [4] ou encore Smice Pilot [5]. Ces outils numériques permettent d'organiser des enquêtes mystères et de produire des rapports visant à montrer les différents axes de dysfonctionnement des services proposés

2. http ://www.converso.com
3. https ://www.orphee.fr/
4. https ://visites-mysteres.ch/
5. https ://smice.com/

par un réseau de vente. Dans cette thèse, nous allons nous intéresser au logiciel Retaily (http ://www.retaily.fr/). Le logiciel Retaily est développé par l'entreprise Effet B qui a travers cette thèse a décidé d'améliorer l'efficacité du traitement de sa grande masse de données collectées durant les enquêtes mystères.

Le déroulement des enquêtes mystères avec le logiciel Retaily (Figure 1 ci-dessous) commence par le paramétrage des enquêtes, c'est-à-dire la définition des compétences à évaluer, la configuration des questionnaires des enquêtes et le choix des points de vente à visiter. Après le paramétrage, le commanditaire des enquêtes lance le recrutement des clients mystères par le biais du logiciel. Ensuite, ce dernier laisse le temps aux clients mystères recrutés de réaliser leurs visites mystères dans les points de vente et de renseigner les questionnaires d'enquêtes. Enfin, les rapports des enquêtes sont générés par le logiciel à la fin des enquêtes et sont transmis à un expert pour dégager les améliorations à apporter dans les points de vente visités.

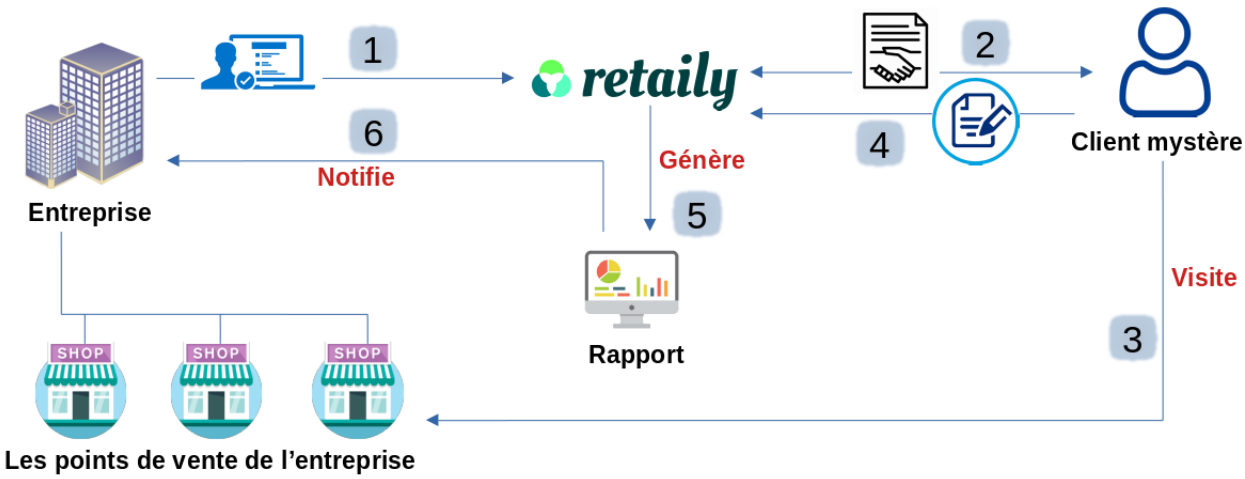

FIGURE 1 – Déroulement d'une enquête mystère avec le logiciel Retaily

Les rapports générés par ces outils de pilotage d'enquêtes mystères dévoilent les statistiques sur les différents services qui ont été évalués, mais ne procurent pas de recommandations pour guider l'expert sur les axes à améliorer au sein des points de vente.

Problématiques et hypothèses de recherche

Avec la numérisation du traitement des données d'enquêtes mystères, les logiciels de pilotage de ces dernières sont particulièrement touchés par le problème de surcharge d'informations, car les résultats obtenus de ces enquêtes renferment un volume d'informations assez conséquent (C.-H. S. Liu et al., 2014 ; Dennis et al., 2001 ; Amudha et al., 2018). De ce fait, les experts en charge de la prise de décisions sont confrontés à plusieurs problèmes : ils sont submergés par le nombre très important d'informations dans l'espace qu'ils explorent. L'exploitation de cette masse d'informations est très complexe pour eux et ils doivent passer beaucoup de temps pour trouver les points à améliorer au sein du réseau de vente qui a sollicité les enquêtes mystères (Rakoto, 2005). De plus, les experts ont des difficultés pour voir ce qu'ils devraient voir ou ce qu'ils pourraient considérer comme important, l'ensemble des résultats d'enquêtes qu'ils évaluent n'est alors en général pas réfléchi, ou bien ils se limitent à voir les items les plus populaires comme dans la plupart des recommandations faites entièrement par l'humain. En conséquence, ils peuvent perdre du temps en regardant des informations ou en explorant des points d'intérêt qui ne les intéressent pas dans leur étude. Inversement, ils peuvent manquer des informations ou des points d'intérêt qui auraient pu les intéresser.

Un des champs de recherche principaux relatifs à la problématique de la surcharge d'information est le domaine de la recherche d'information (Hwang & Lin, 1999 ; N. Davis, 2011 ; Bawden et al., 1999). Le principe général est d'élaborer des méthodes et des algorithmes afin de rechercher des ressources (par exemple, des pages web, des films et dans notre cadre d'application des œuvres ou des points d'intérêt) en fonction de requêtes formulées par des utilisateurs. Il n'est cependant pas toujours évident pour un utilisateur de savoir comment exprimer sa demande. De plus, sa requête correspond généralement à une quantité importante de ressources et il est difficile de savoir quels résultats lui présenter en premier, d'autant plus que d'un utilisateur à un autre, l'ordre de priorité peut changer.

Un autre champ de recherche relatif à cette problématique est le domaine des systèmes de recommandation (Aljukhadar et al., 2012 ; Lu et al., 2012 ; Costa & Macedo, 2013). Ces systèmes sont capables de fournir des recommandations adaptées aux préférences et aux besoins des utilisateurs. Ils se sont avérés être très satisfaisants pour aider les

utilisateurs à accéder aux ressources désirées dans un temps limité. Initialement conçus pour la recommandation de ressources web, films,etc. les systèmes de recommandation sont devenus de plus en plus populaires et sont aujourd'hui un composant principal de beaucoup d'applications dans différents domaines (Lu et al., 2012 ; Farzan & Brusilovsky, 2011).

Nos travaux sont focalisés sur les systèmes de recommandation, notamment sur les propositions automatiques de plans d'action pour booster les forces de vente d'un réseau de vente. Nos recherches sont appliquées sur le logiciel Retaily, une plateforme numérique de pilotage d'enquêtes mystères.

L'application Retaily recueille les données subjectives des clients mystères via un questionnaire. Après traitement des données renseignées par les clients, le logiciel Retaily génère un rapport. Ce rapport représente les résultats des enquêtes effectuées et sont décrites sous forme de diagrammes statistiques (Figure 1). La faiblesse principale de cette présentation concerne leur interprétation : aucun jugement argumenté n'est formulé, aucune évaluation qualitative des résultats n'est disponible et aucune solution pour améliorer les scores constatés n'est proposée. En synthèse, la présentation de ces résultats exige un temps de travail interprétatif supplémentaire de la part d'un expert afin de :

- réaliser un diagnostic holistique (qui prenne en compte toutes les interdépendances entre les dimensions de l'évaluation) et analytique (qui se focalise sur chaque dimension) ;
- proposer des suggestions pertinentes sur les changements à apporter afin d'améliorer la qualité et l'efficacité du service évalué.

Ainsi nos questions de recherche sont :

1. Est-il possible d'automatiser partiellement le diagnostic et l'analyse des résultats issues des enquêtes mystère ?
2. Comment répondre au verrou d'hétérogénéité des données du système ?
3. Comment lever le verrou d'hétérogénéité d'usage sur les aspects adaptatif du système ?
4. Peut-on tirer de ce diagnostic une première liste de recommandations primaires pertinentes ?

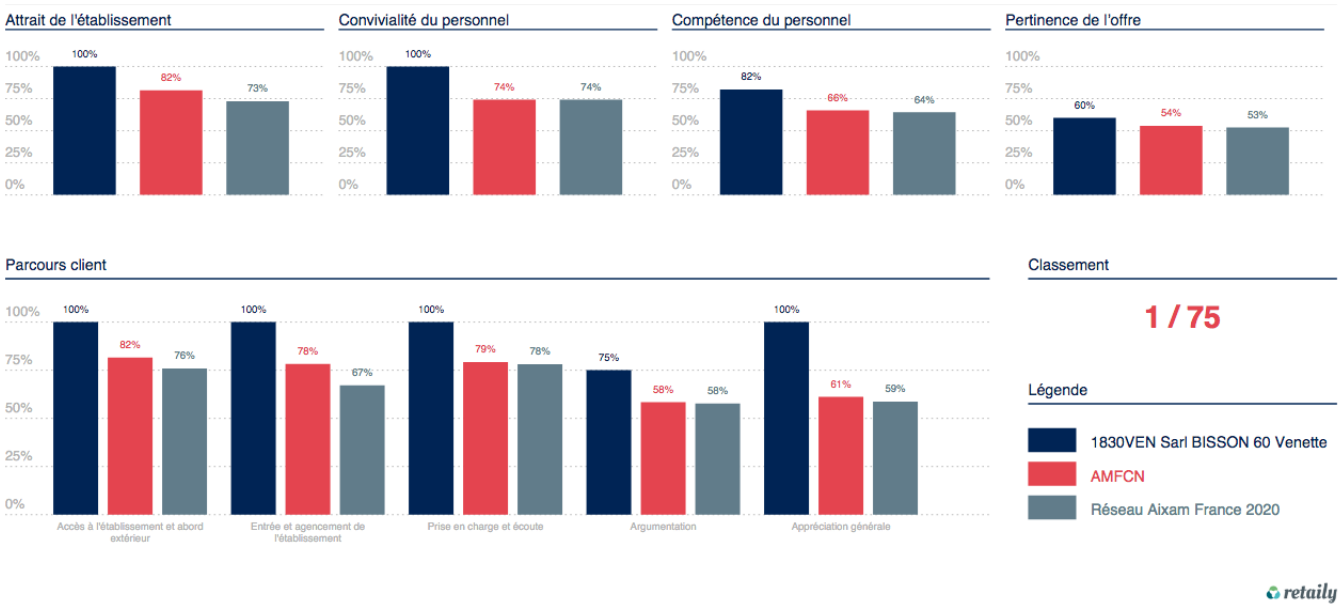

FIGURE 2 – Rapport d'enquête mystère (www.retaily.fr).

Méthodologie

Tout travail de recherche est basé sur une représentation du monde, utilise une méthodologie, justifie des résultats permettant d'expliquer et de faire comprendre. Une explicitation de ces suppositions épistémologiques permet de gérer la démarche de recherche, de faire évoluer le niveau de la connaissance qui en est issue (Allard-Poesi et al., 2014, 1999).

L'étude des systèmes d'information appliquée au marketing participe à l'amélioration de la performance commerciale (Cron & Sobol, 1983). Les systèmes d'information ont une influence sur la réussite d'une stratégie en marketing. Les travaux de Henderson et al. (1992) ont traité la relation entre système d'information et stratégie en marketing et ont montré qu'une stratégie marketing efficace peut apporter un soutien aux entreprises de commerce.

Les travaux que nous avons menés dans cette thèse s'inscrivent dans les théories des systèmes d'information, de l'ingénierie des connaissances et du marketing au sein du domaine de l'informatique et des Sciences de l'Information et de la Communication (SIC). Nos travaux portent particulièrement sur le filtrage d'information, notamment sur les systèmes de recommandation pour proposer automatiquement aux entreprises de com-

merce des stratégies marketing. Les travaux existants (Burke, 2002 ; Resnik, 1995 ; Kunaver & Požrl, 2017 ; Schafer et al., 2007 ; Schein et al., 2002 ; Nguyen et al., 2006 ; Safoury & Salah, 2013 ; Berrichi & Djouaher, 2020) ne nous permettent pas de lever les verrous cités précédemment dans la section problématiques et hypothèse de recherche. Nous considérons que les technologies sémantiques et celles de l'apprentissage automatique permettraient une réduction des temps d'analyse et de prise de décision, notamment en :

— automatisant partiellement l'expertise humaine ;
— levant le verrou d'hétérogénéité des données ;
— améliorant le démarrage à froid, dans le cas d'un nouvel utilisateur dans le système ;
— améliorant aussi la pertinence des prédictions pour la recommandation ;
— réalisant une bonne classification des items recommandés.

Pour atteindre ces objectifs, il est nécessaire d'évaluer l'opportunité d'extraire, décrire, modéliser l'expertise d'un agent humain en vue de l'implémenter dans un système de recommandation sémantique capable d'interpréter des données issues d'enquêtes mystère. En effet, l'enjeu est de répondre à la fois au verrou d'hétérogénéité sémantique (système statique) et au verrou d'hétérogénéité d'usages (dynamique et adaptatif du système). A ce titre, notre démarche de résolution sera effectuée en deux parties.

Une première partie (base de connaissances) où les données et les processus métier sont modélisés par une ontologie et des règles métier (agissant sur le modèle de plans d'action, le modèle de contraintes et le modèle de contexte).

La deuxième partie représentera les modèles adaptatifs et sera basée sur des algorithmes reproduisant les heuristiques des experts et de l'ontologie pour aligner les contenus du domaine du marketing.

L'accroissement du nombre de sources de données produites par Effet B et la quantité d'informations (le Big data) gérées au sein de ces sources nécessitent la mise en place d'un système capable d'extraire automatiquement les connaissances directement disponibles ou dissimulées dans la complexité des données induite par leur hétérogénéité et leur accumulation exponentielle. Les technologies de représentations de bases de connaissances permettent la mise au point de tels systèmes, car elles décrivent les données et raisonnent en s'appuyant sur des aspects du monde réel.

Contributions de la thèse

Dans cette thèse, nous proposons un système de recommandation qui est construit sur la base d'une combinaison de plusieurs technologies. Ces technologies sont celles du web sémantique et celles de l'apprentissage automatique. Notre système est basé sur trois modules : un premier module de préparation et représentation des connaissances (technologies sémantiques), un deuxième module de modélisation et de prédiction pour la recommandation (système de recommandation et apprentissage automatique) et enfin un dernier module qui est à la charge de la classification des items obtenus (algorithme de classification). Les contributions apportées dans cette thèse sont :

1. un prototype d'automatisation partielle de l'analyse de l'expertise humaine ;
2. une méthode d'homogénéisation des données ;
3. une méthode d'amélioration du démarrage à froid pour un nouvel utilisateur ;
4. une méthode d'amélioration de la prédiction dans un processus de recommandation ;
5. une méthode de classification des items recommandés.

Contexte de la thèse

Le sujet de thèse s'inscrit dans la continuité des travaux de recherche et de développement menés depuis 2014 par l'entreprise EFFET B [6] sur le logiciel Retaily. Retaily est un logiciel en ligne d'organisation d'enquêtes réalisées par des clients mystère, ainsi que d'études de satisfaction.

Ce logiciel est un outil qui propose instantanément des questionnaires, scénarios et modèles de rapports d'enquêtes, à partir de modèles interactifs et sur une même interface. EFFET B est soucieuse de conserver l'avance conceptuelle et technologique de Retaily, et souhaite automatiser certains aspects du processus d'analyse des résultats de l'enquête en exploitant les avancées du Web Sémantique, domaine d'expertise du laboratoire ELLIADD [7]

La recherche doctorale qui est présenté dans cette thèse a été menée au sein du pôle Conception Création Médiations (CCM) du laboratoire ELLIADD. Le pôle CCM réunit autour de la problématique de la médiation une équipe de chercheur.e.s pluridisciplinaires issues des sciences humaines et sociales, dont les Sciences de l'Information et de la Communication (SIC), les sciences de l'éducation et les sciences du langage. Les recherches menées dans le cadre du pôle CCM s'articulent autour de plusieurs programmes scientifiques, parmi lesquels figurent : *Sémantisation des contenus et la représentation des connaissances*.

J'ai rejoint le pôle CCM en janvier 2017 dans le cadre de cette thèse dont le sujet a été construit durant mon stage de fin d'étude de Master en Produits et Service Multimédia (PSM). Ce stage a été réalisé au sein de l'entreprise Effet B pour une durée de six mois. Au terme de ce stage en juillet 2016, nous avons soumis auprès de l'Association Nationale Recherche Technologie (ANRT) notre projet de recherche pour un financement par convention CIFRE [8] et une réponse positive a été émise à notre demande en décembre 2016.

6. Créé en 2009, EFFET B est un studio de création de sites internet à Lyon qui aide ses clients dans l'aboutissement de leurs projets de communication sur le web.

7. L'EA 4661 ELLIADD (Edition, Littératures, Langages, Informatique, Arts, Didactiques, Discours) est une unité de recherche de l'Université de Franche-Comté reconnue par le Ministère et évaluée A par l'AERES pour son projet scientifique.

8. Convention Industrielle de Formation par la Recherche, est un dispositif qui subventionne toute entreprise de droit français qui embauche un doctorant pour le placer au cœur d'une collaboration de recherche avec un laboratoire public. Ce dispositif est créé en 1981 et est géré par l'ANRT.

Ainsi, cette thèse s'inscrit dans le cadre d'un contrat CIFRE entre l'entreprise EFFET B et le laboratoire ELLIADD de l'université Bourgogne Franche-comté.
Pendant la thèse, j'ai mené mes travaux principalement de manière individuelle, hormis les séjours passés au laboratoire ELLIADD pour échanger de vive voix avec mes directeurs de thèse et pour assister aux manifestations scientifiques.

Plan du manuscrit

Le manuscrit de la thèse est structuré en quatre chapitres présentant les travaux effectués au cours de la recherche doctorale.

Le premier chapitre dresse un état de l'art sur les systèmes de recommandation. La recommandation est un processus visant à proposer des items susceptibles d'intéresser une personne ou un groupe de personnes. Ce processus tient compte des notes antérieures de chaque profil et leur historique d'achat ou d'intérêt (Burke, 2002).
Dans un premier temps, nous présentons ici les caractéristiques des différents types systèmes de recommandations, faire un état de l'art de ces derniers en détails et montrer leurs impacts sur les plateformes de e-commerce. Dans un second temps, nous y abordons les méthodes de recommandation basées sur le contenu, ces méthodes de recommandation sont divisées en deux groupes de méthodes : un premier groupe de méthodes recommandations traditionnelles et un deuxième groupe de méthodes récentes utilisant les technologies de l'apprentissage automatique. La troisième partie présente la recommandation basée sur le filtrage collaboratif. En quatrième partie, les méthodes de recommandation hybride qui regroupe deux ou plusieurs méthodes de recommandations traditionnelles ou récentes. En cinquième partie, nous présentons les autres méthodes de recommandation utilisées dans des champs de recherche bien précis. En dernier et sixième partie les avantages et inconvénients de méthodes de recommandation seront énumérés. L'accent sera mis sur le type filtrage collaboratif qui est un type de recommandation qui va plus nous intéresser dans cette thèse, car nous avons focalisé nos recherches sur la similarité sémantique entre les profils des points vente pour produire des recommandations.

Le second chapitre se concentre sur l'expertise humaine et la représentation des connaissances. Selon Fornel (1990, p. 65), « *l'expertise humaine est une compétence située qui, en tant qu'elle suppose une forme d'enquête pratique, émerge comme une propriété des comportements que l'on observe quand les individus réalisent des activités* ». L'objectif de l'expertise est de fournir des connaissances, mais ceci ne signifie pas que l'expertise puisse se définir purement et simplement comme l'expression d'une connaissance (Roqueplo, 1997, p. 14). La représentation de ces connaissances sont en partie constituées d'un ensemble de capacités qui est requis pour fournir la solution des problèmes auxquels l'expertise s'applique (Fornel, 1990, p. 65).
Dans un premier temps, l'accent est mis sur les concepts d'expertise humaine et un état de l'art sur les méthodes et modèles d'analyse de l'expertise humaine est réalisé. Dans un second temps, la représentation des connaissances de l'expertise humaine sera abordée dans un contexte de recommandation. Nous allons voir comment les connaissances de l'expertise humaines sont organisées en utilisant les technologies du web sémantique.

Le troisième chapitre traite la modélisation du système de recommandation sémantique enrichi que nous proposons dans cette thèse. Ce chapitre aborde nos différentes contributions que nous apportons à travers notre système de recommandation.
Dans un premier temps, la présentation de l'architecture globale de notre système de recommandation sémantique enrichi est effectuée. Les différents modules qui composent notre architecture seront modélisés.
L'objectif de ce chapitre est de présenter la modélisation de notre méthode combinatoire de recommandation sémantique enrichie. L'implémentation est une analyse empirique et des évaluations du rendement seront effectués dans le chapitre suivant.

Le quatrième chapitre détails les implémentations et les évaluations qui ont été effectuées pour la validation de nos différentes hypothèses. Les résultats de ces évaluations seront présentées avec des données issues d'enquêtes mystère. Enfin, une discussion sera faite et une synthèse sur la validation des différentes contributions qui ont été apportées dans cette thèse sera effectuée.

Nous concluons en faisant le bilan des recherches qui ont été effectuées et en présentant les perspectives dont elles résultent. Nous présentons les différents verrous qui ont été levés par les contributions qui ont été proposées dans cette thèse.

Table des matières

Chapitre 1

Système de recommandation

Sommaire

« Il est souvent nécessaire de prendre une décision sur
la base de connaissances suffisantes pour l'action
mais insuffisantes pour satisfaire l'intellect. »
Emmanuel Kant, Artiste, écrivain, Philosophe (1724 - 1804)

Durant deux décennies, l'ordinateur a progressivement envahi tous les domaines de l'activité humaine et a permis de numériser l'information. Avec la création du Web en 1989, les réseaux internet et intranet sont devenus la structure centrale des systèmes informatiques des entreprises, notamment celles commercialisant des produits. La numérisation du processus de vente de ces entreprises a conduit à une augmentation considérable des collections de données en produisant trop de résultats à travers les requêtes lancées.
Face à la quantité grandissante de l'offre de produits et services proposés dans le secteur du commerce en ligne, les consommateurs apprécient de plus en plus les aides à la prise de décision qui leur sont suggérées lors de la phase d'achat. Ces aides, nommés systèmes de recommandation (SR) sont un enjeu majeur pour les plateformes de e-commerce notamment pour construire la confiance auprès des consommateurs.
De nombreuses méthodes de recommandation ont été élaborées (des méthodes traditionnelles et récentes). Dans cette thèse nous considérons comme méthodes traditionnelles de recommandation celles utilisant des technologies n'appartenant pas à l'apprentissage automatique et méthodes récentes celles qui les utilisent. Parmi les plus récentes figure l'apprentissage automatique qui est au cœur du SR que nous présentons dans cette thèse.
Dans ce premier chapitre, nous présentons les origines des SR et leurs domaines d'application, leurs typologies et les méthodes de conception existantes, en nous appuyant sur des exemples issus de notre terrain d'étude, à savoir le e-marketing. Nous montrons également les avantages et inconvénients des SRs en nous focalisant sur deux types spécifiques de SR : le filtrage collaboratif et le filtrage sur le contenu.

1.1 Origines et applications des SR

La recommandation peut être comparée à un dialogue entre une personne experte d'un domaine et une autre désireuse d'améliorer sa propre connaissance. Par exemple, un

bibliothécaire peut recommander un ensemble de livres à l'un de ses clients sur la base de l'historique d'achats ou d'emprunts de ce dernier.
(Burke, 2002) définit un SR comme un système informatique capable de fournir des recommandations qui proposent à l'utilisateur des ressources pertinentes pour répondre à ses besoins. Les SR évaluent les préférences d'un utilisateur pour proposer des ressources pertinentes. Ils ont pour but de rendre facile le traitement des informations dans une grande quantité de données.

1.1.1 Origines des SR

La naissance des SR remonte aux années 90 avec le système Tapestry, conçu par D. Goldberg et al. (1992), au Palo Alto Research Center, ils ont inventé l'expression « filtrage collaboratif » (FColl) que nous avons expliqué à la page 31 de ce chapitre. La préférence du terme général de « SR » a été adoptée pour deux raisons :

1. les commanditaires ne sont pas en mesure d'expliquer leur collaboration entre les bénéficiaires des recommandations, car ils ne se connaissent probablement pas ;
2. les recommandations peuvent suggérer des éléments particulièrement intéressants, en plus d'indiquer ceux qui devraient être filtrés (Resnick & Varian, 1997).

D'autres SR ont vu le jour en 1994 et en 1995, tels que le SR de films développés par GroupLens [1] (Resnick et al., 1994) et le SR de musique Ringo [2] proposé par Shardanand et Maes (1995). Ces deux systèmes sont également basés sur le FColl. Quelques années plus tard, avec le développement de l'Internet et des applications web, les SR ont évolué et ont été implémentés dans différents domaines d'application. Parmi ces SR, nous pouvons citer ceux présentés dans le Tableau 1.1.

1. Un laboratoire de recherche qui travaille explicitement sur le problème de recommandation automatique dans le cadre des forums de news. https ://grouplens.org/
2. Un système de recommandation de musique, basé sur les appréciations des utilisateurs, il a été créé en 1995.

Domaine d'applications	SR	Références
Distribution de Films	Eachmovie	Breese et al. (1998)
Distribution de Films	Movielens	Harper et Konstan (2015)
Recrutement	Job-Finder	Luu et Vasavda (2020)
Divertissement	Jester	K. Goldberg et al. (2001)
Bibliographie	Citations bibliographiques	McNee et al. (2002)
Restauration	SR de restaurations	Burke (2002)
e-commerce	SR d'Amazon	Linden et al. (2003)
Recherche d'information	Le moteur de recherche d'AOL	Vernette et Marketing (2007)
Musique	LastFM	Levy et Bosteels (2010)

TABLE 1.1 – Quelques exemples de SR

1.1.2 Applications des SR

Le nombre croissant des produits disponibles sur les sites e-commerce a rendu le choix des consommateurs très complexe (Isaac & Volle, 2014). Pour ces sites de vente en ligne, un SR est considéré comme un outil incontournable pour leur stratégie de marketing. En effet, ces dernières années, beaucoup de sites web de vente en ligne utilisent les SR pour booster leur vente. Ce domaine d'étude est devenu un champ de recherche important. Amazon est connu pour l'utilisation d'un SR fiable (Linden et al., 2003). Cette pratique se généralise aujourd'hui et touche même les petits sites e-commerce. Les méthodes de recommandations les plus utilisées par ces sites de vente sont : la recommandation d'objet, la recommandation sociale et la recommandation hybride. Ces méthodes sont fondées sur deux types principaux de SR qui sont le FColl et le filtrage du contenu (FCont).

Recommandation d'objet

La recommandation d'objet est une méthode qui se focalise sur le contenu d'une plateforme sur laquelle la recommandation a été sollicitée. Elle est dans la famille du type de recommandation basé FCont. Elle s'appuie, par exemple, sur les caractéristiques des produits comme les couleurs, les tailles, les marques. Elle se base aussi sur l'historique de l'utilisateur. La Figure 1.1 illustre la recommandation d'objet.

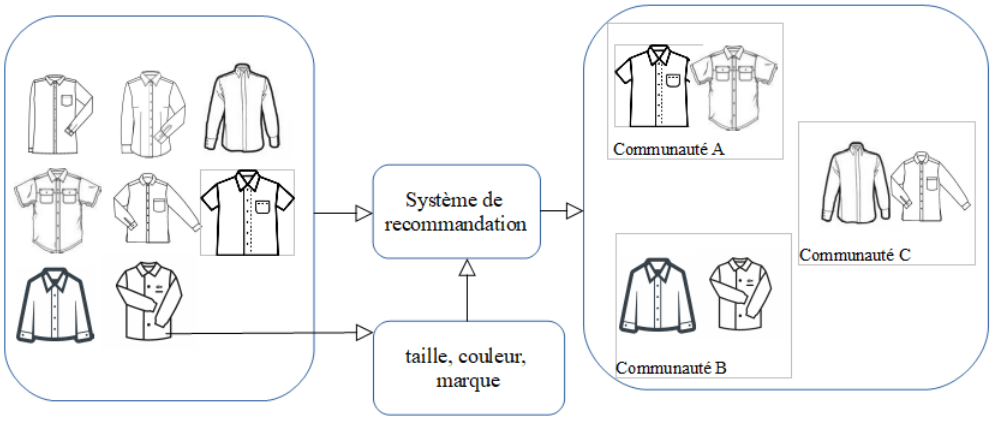

FIGURE 1.1 – Recommandation Objet

Recommandation sociale

La recommandation sociale est une méthode de recommandation basée sur le FColl. Le processus de ce type de recommandation est le suivant : quand un consommateur X achète les produits F et D, le consommateur Y qui achète le produit F est supposé être intéressé par le produit D. Les consommateurs X et Y sont considérés comme de proches voisins dans un modèle de données. Ce modèle de données est construit par le système à partir des données collectées. La recommandation sociale prend en compte les intérêts de l'ensemble des consommateurs, ce qui la différencie de la recommandation objet comme le montre la Figure 1.2 ci-dessous.

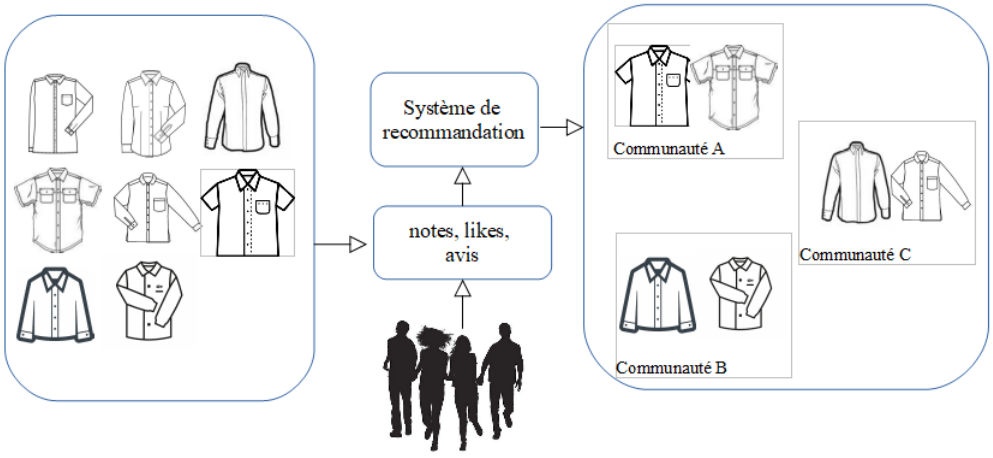

FIGURE 1.2 – Recommandation Sociale

Recommandation hybride

La recommandation hybride est utilisée sur les sites de vente en ligne, la recommandation hybride est la combinaison de la recommandation objet et la recommandation sociale. Elle est représentée par la Figure 1.3 ci-dessous.

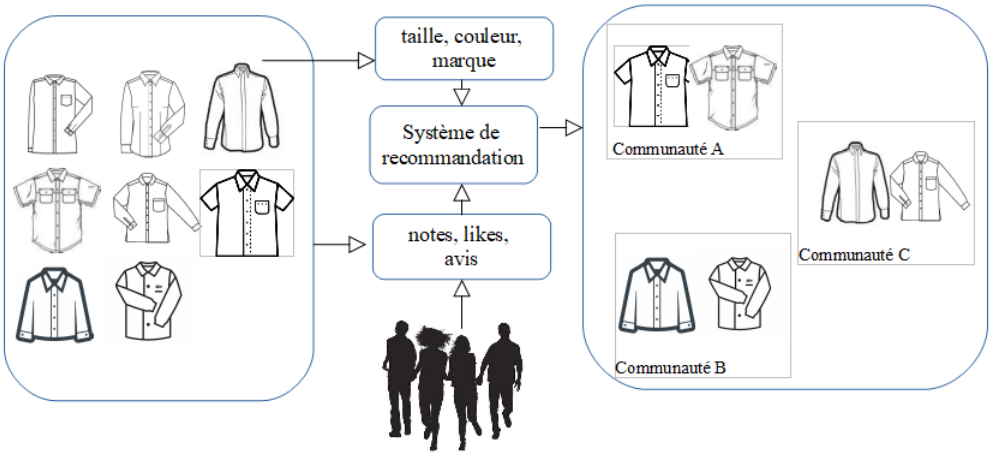

FIGURE 1.3 – Recommandation hybride

Quelques SR de grandes entreprises

Amazon

Amazon[3] possède un SR très performant, il utilise trois méthodes (Linden et al., 2003) :

1. la recommandation basée des objets sur s'appuie sur le comportement passé de l'utilisateur, ce qu'on peut assimiler à une recommandation personnalisée. Le comportement de l'utilisateur est prédit à partir de son historique de navigation et de son historique d'achat ;
2. la recommandation sociale basée sur le FColl. Cette recommandation est fondée sur les comportements des autres utilisateurs ;
3. la recommandation objet, qui utilise les caractéristiques de l'objet (recommandation basée sur FCont) pour faire des recommandations.

3. Amazon est une entreprise américaine de commerce électronique, créée par Jeff Bezos en juillet 1994

Lorsqu'on fait un achat sur Amazon, on a l'habitude de lire le message d'Amazon suivant : « les personnes qui ont acheté ou regardé le produit x ont aussi acheté le produit y ». Cette phrase est issue d'une recommandation basée sur la méthode du « plus proche voisin » que nous allons voir en détail plus bas dans ce chapitre. C'est une méthode de recommandation sociale basée sur FColl. L'image ci-dessous, montre les produits susceptibles d'intéresser l'utilisateur car ce dernier possède un historique d'achat avec d'autres utilisateurs qui ont acheté les produits que Amazon lui recommande.

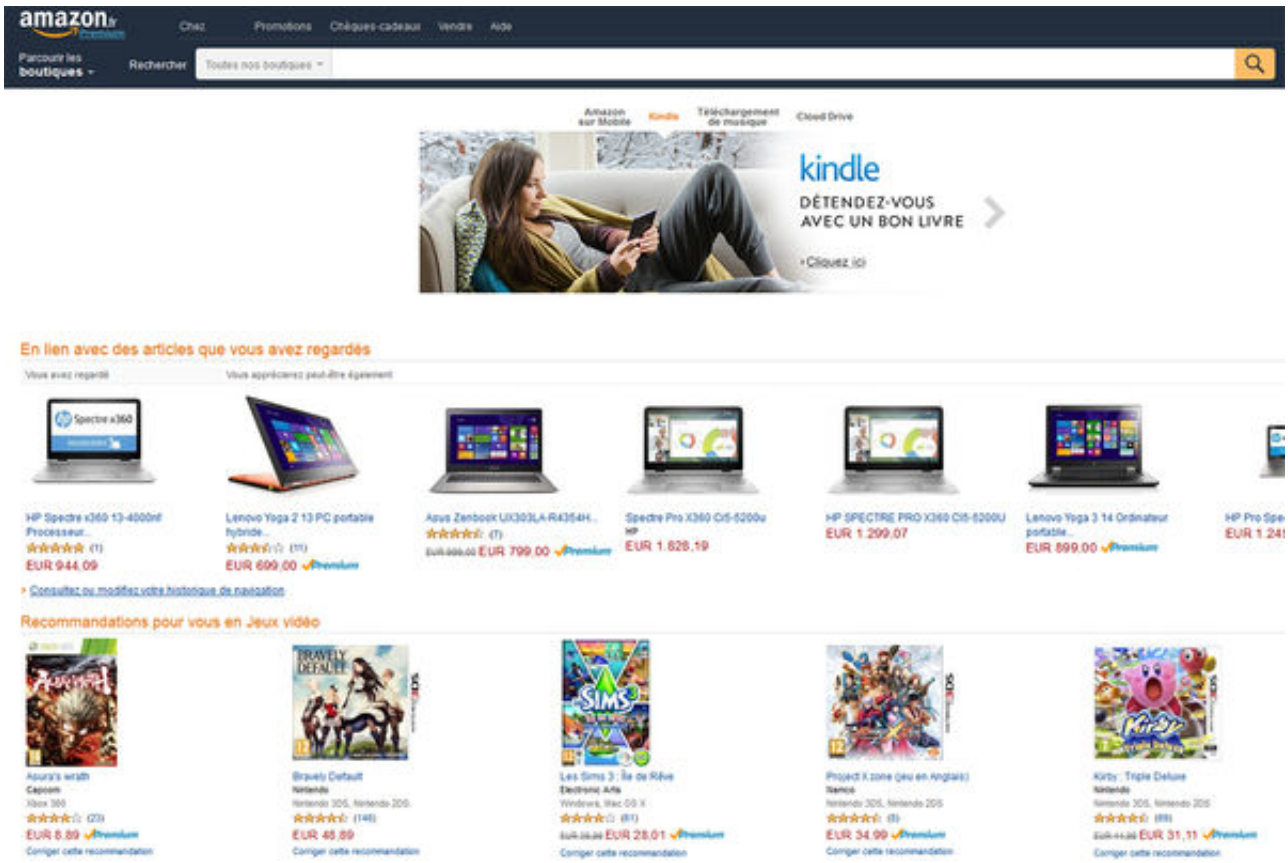

FIGURE 1.4 – Recommandations d'Amazon

Netflix

Le SR de Netflix [4] a pour objectif d'aider ses abonnés à trouver facilement une série télé ou un film qui leur plaira dans son catalogue. La probabilité qu'un utilisateur visionne un titre particulier du catalogue est évaluée en fonction d'une liste de facteurs :

- les interactions de l'utilisateur avec le service (e.g. la nivagation, les cliques, les visionnages) ;
- les utilisateurs possédant des goûts similaires ;
- les informations sur les films et les séries (e.g. le titre, la catégorie, les acteurs, le genre).

Pour mieux personnaliser ses recommandations, Netflix prend aussi en compte :

- le moment de la journée où l'utilisateur visionne les contenus ;
- les appareils qui ont été utilisés pour visionner et la durée du visionnement.

Dans le processus de prise de décision, le SR de Netflix ne tient pas compte des aspects démographiques, il suggère des recommandations basées sur les habitudes de visionnages de contenus d'utilisateurs similaires, ce qui est une recommandation sociale ou encore recommandation basée sur le FColl. Il propose aussi des contenus qui partagent des caractéristiques avec des films que l'utilisateur a noté de manière positive, ce que l'on peut considérer comme une recommandation objet (recommandation basée sur FCont). La Figure 1.5 est une illustration du processus de recommandation de Netflix.

FIGURE 1.5 – Recommandation Netflix

4. Netflix est une entreprise américaine de distribution et d'exploitation d'œuvres cinématographiques et télévisuelles créée en 1997 par Reed Hasting et Marc Randolph. Chaque client peut, après avoir visionné un film, donner son avis sur ce dernier. https ://www.netflix.com/fr/

En plus de recommander l'utilisateur sur sa page d'accueil de visionnage, Netflix fait un classement des titres et les catégorise. Le processus de classement et de catégorisation est réalisé en utilisant des algorithmes et des systèmes complexes, tels que ceux de l'apprentissage automatique (Machine Learning). Ces algorithmes complexes permettent de personnaliser l'expérience de l'utilisateur.

Le SR de Netflix est une bonne illustration de la recommandation hybride. En 2006, Netflix a lancé le prix Netflix dans le but d'améliorer les méthodes traditionnelles de la recommandation de contenus (Bennett et al., 2007). Avant la compétition, il utilisait un SR, « CineMatch », permettant de proposer aux clients une sélection de films ou de séries. En réalisant qu'un bon processus de recommandation était un moyen pertinent de fidéliser sa clientèle et d'augmenter son chiffre d'affaires, Netflix a cherché à apporter une amélioration à son moteur de recommandation. Le but de la compétition était de mettre en place un SR meilleur que « CineMatch » dans les tests. Cette compétition a suscité l'intérêt chez les amateurs de films, mais plus encore dans le monde de la recherche scientifique. Netflix avait promis un million de dollars au vainqueur. Après trois ans de compétition, le prix a été remporté par l'équipe Bellkor's Pramatic Chaos. L'équipe a proposé une solution qui fait appel à une hybridation de plus de cent modèles. Cette technique d'hybridation a été abordée dans plusieurs articles (Piotte & Chabbert, 2009 ; Koren, 2009). Les solutions proposées par l'équipe vainqueur étaient pertinentes, mais gourmandes en termes de calcul et en mémoire. La compétition a cependant permis de mettre en évidence l'atout des méthodes de factorisation pour la résolution de problématiques liées à la recommandation avec l'utilisation d'informations complémentaires, comme les effets temporels, les niveaux de confiance et les évaluations implicites (Bell & Koren, 2007). Les méthodes de factorisations permettent de simplifier les expressions mathématiques afin de résoudre un problème plus simplement (Champagne, 2004).

Google

Google, quant à lui, se focalise sur la combinaison de trois méthodes pour améliorer son moteur de recherche :

— la recommandation sociale, en faisant appel à l'algorithme du PageRank [5] qui utilise les liens entre les pages web. La recommandation sociale est exploitée sur les contenus provenant des communautés de Google+ ;

— Google personnalise nos résultats de recherche en se basant sur notre géolocalisation et nos dernières recherches. Quand on se connecte à Google par exemple, il propose un contenu encore plus pertinent sur la base de notre historique de recherche, ce qui est une recommandation personnalisée ;

— la recommandation objet ou encore basée sur le contenu est utilisée par Google dans une approche sémantique pour sa fonction *Did you mean* [6].

L'API Google Maps de Google est utilisée dans les SR basées sur la géolocalisation, on peut citer les travaux de Benouaret (2017) sur la recommandation contextuelle et composite pour la visite personnalisée de sites culturels. L'idée de ses travaux est de créer un système permettant d'améliorer l'expérience des visiteurs de musées en leur recommandant les œuvres qui correspondent à leurs préférences et qui sont susceptibles de les intéresser. Pour ce faire, Benouaret (2017) a utilisé un système hybride qui combine trois méthodes différentes de recommandation : démographique, sémantique et collaborative de manière séquentielle en fonction de la progression de la visite. Nous avons expliqué ces méthodes de recommandation à la page 38 de ce chapitre.

Ainsi, en fonction des préférences de l'utilisateur et à partir des résultats obtenus par ses trois méthodes, le système de Benouaret (2017) utilise l'API de Google pour le rendu final des emplacements recommandés pour l'utilisateur (Benouaret, 2017).

5. PageRank est un algorithme d'analyse de lien qui attribut une pondération numérique à chaque élément d'un ensemble de documents comportant des hyperliens. Il peut être appliqué à n'importe qu'elle collection d'entités avec des citations et des références réciproques. Son classement résulte d'un algorithme mathématique basé sur le graphique Web, créé par toutes les pages web.

6. *Did you mean* est un algorithme de Google qui suggère à l'utilisateur une autre orthographe des mots clés qui sont recherchés via le moteur de recherche.

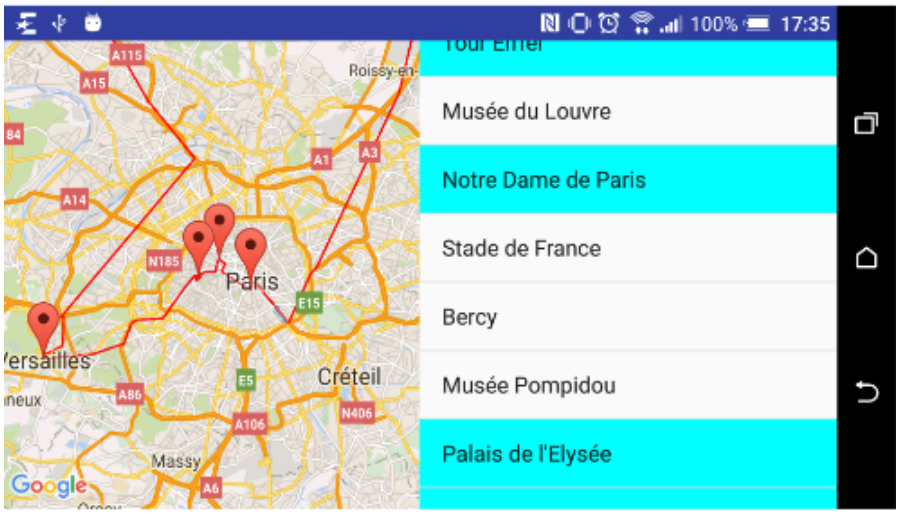

FIGURE 1.6 – Recommandation avec Google Map

1.2 Concepts d'un SR

En science de l'information, un SR est un outil qui permet aux utilisateurs d'exprimer leurs préférences sur différents items (Kembellec et al., 2014). Un SR s'appuie sur les entités utilisateur, item et sur la notion de note.

1.2.1 Entité utilisateur

L'entité utilisateur varie selon une logique d'échelle et de contexte d'application, l'entité utilisateur peut être représentée par un individu, par un groupe d'individus ou une organisation, ou un composant logiciel. Par exemple, la plateforme Netflix qualifie l'utilisateur comme une personne unique pendant le processus de recommandation (Gomez-Uribe & Hunt, 2016).

En revanche, dans le cadre de notre thèse, l'entité utilisateur peut être le client qui a demandé le sondage (une organisation), le professionnel ayant les compétences métiers qui font l'objet de ce sondage (i.e. e.g le graphiste, le développeur, etc.) ou bien un module logiciel responsable d'un traitement des données. Nous illustrerons ce troisième cas dans le chapitre 3 décrivant l'architecture de notre système.

Dans notre projet, nous avons conceptualisé un groupe d'utilisateurs en termes de communauté. D'après Perugini et al. (2004), les communautés sont formées sur la base de critères comme la similarité sur les évaluations faites par les utilisateurs. Selon Nguyen et

al. (2006), on peut s'appuyer sur la formation de communautés sur de multiples critères considérés significatifs pour le système par exemple la géolocalisation pour un réseau de points de vente. Les communautés sont étudiées sous un aspect social dans un objectif d'établir des relations entre les utilisateurs (Perugini et al., 2004). Pour Montaner et al. (2003) ces communautés sont aussi étudiées dans un aspect fonctionnel. Des travaux existants montrent que les communautés sont formées sur un unique critère (Nguyen et al., 2006), On rencontre généralement dans les systèmes, un ensemble hétérogène de critères sur lesquels s'appuie la formation de communautés. Nous montrerons dans le chapitre 3 comment nous avons instancié ce concept de communauté d'utilisateurs.
Les SR forment des communautés d'utilisateurs en fonction de leur historique d'utilisation des services proposés par le site. En plus de la formation des communautés, les SR s'appuient sur les caractéristiques connues de l'utilisateur (e.g. sexe, âge, secteur d'activité) ou sur une combinaison de ces caractéristiques et de son historique. Ainsi, pour faire une recommandation, le SR va d'abord chercher la communauté à laquelle l'utilisateur appartient afin de lui proposer des offres susceptibles de l'intéresser.

1.2.2 Entité item

L'entité item est une description d'un ensemble d'attributs, selon une représentation structurée des données du SR, telle qu'un modèle vectoriel (Baloian et al., 2004). Cette représentation permet de décrire la relation que détient l'item avec les utilisateurs. Ce modèle vectoriel est un vecteur de poids, où chaque poids est associé à un terme. Un poids est une valeur numérique correspondant à la présence ou à la fréquence ou encore à l'importance du terme dans l'item.
Un attribut peut être de deux types : numérique ou catégoriel. Les attributs de type numérique peuvent être continus, par exemple, la taille, le poids ou la durée d'un événement, ou discrètes, comme le nombre de produits commandés sur un site e-commerce. Il est possible de transformer des données continues en des données discrètes par discrétisation.
Les attributs de type catégoriel sont un ensemble fini de valeurs alphanumériques, par exemple le numéro de passeport ou encore l'évaluation d'un film. Si les attributs catégoriels peuvent être classés, il s'agit d'attributs catégoriels ordinaux, comme dans le cas de

l'évaluation d'un film. Dans le cas contraire, l'attribut catégoriel est nominal (Kassab, 2009).

Le Tableau représente l'ensemble des attributs d'une liste structurée d'items.

Id	Titre de film	Genre	Langue
1	Le roi lion	Drame/Aventure	Anglais
2	La source	Comédie	Français
3	So long, my song	Drame/Famille	Anglais

TABLE 1.2 – Exemple d'une liste structurée d'items

Un item est représenté sous forme de modèle vectoriel, construit à partir de sa représentation sémantique. Si on prend le cas d'une représentation structurée après une discrétisation, un terme est associé à la valeur d'un attribut, le poids est représenté par une valeur booléenne signifiant la présence ou pas de la valeur de l'attribut dans l'item. Si on reprend l'exemple du Tableau 1.2, un ensemble de films est décrit en fonction de trois attributs (le titre du film, le genre et la langue) et un identifiant. Supposons que les valeurs de l'attribut « titre du film » sont : Le roi lion ; La source et So long. Le Tableau 1.3 montre la représentation en modèle vectoriel des items du Tableau 1.2.

Id	Le roi lion	La source	So long	Drame	Aventure	Comédie	Famille	Anglais	Français
1	1	0	0	1	1	0	0	1	0
2	0	1	0	0	0	1	0	0	1
3	0	0	1	1	0	0	1	1	0

TABLE 1.3 – Représentation en modèle vectoriel des items du tableau

La valeur d'un attribut peut être caractérisée de descripteurs, ainsi dans notre exemple « Aventure » est un descripteur. On peut rencontrer des attributs dont le poids représente la fréquence d'un descripteur dans un item comme le cas des annotations. Le poids d'une annotation représente le nombre de fois qu'elle a été employée pour l'annotation de l'item. Dans ce contexte, le poids est défini par l'expression de fréquence suivante :

Soit n : nombre de fréquences de d_i dans i et D_i, l'ensemble des descripteurs

$$frequence_i(d_i) = \begin{cases} n \;\; si \;\; n \in N \\ 0 \;\; si \;\; d_i \in D_i \end{cases} \tag{1.1}$$

Dans le contexte d'un ensemble d'items avec une représentation non structurée, ces derniers doivent subir une transformation pour obtenir une représentation structurée. Cette structuration passe par une indexation de l'ensemble d'items qui va être utilisée dans le filtrage d'informations pour le traitement des documents (Salton, 1989). L'indexation est une opération qui consiste à faire une extraction des mots les plus pertinents contenus dans un document. Après l'étape d'indexation, chaque item est modélisé en une représentation vectorielle. Dans cette représentation, l'item sera décrit par un vecteur de poids dans lequel chaque poids correspond à un mot.
Les informations qui permettent de relier l'entité utilisateur à l'entité item sont de natures différentes : notes, achats, clics, historiques, etc. Les SR se focalisent majoritairement sur l'utilisation de « note » (Adomavicius & Tuzhilin, 2005).

1.2.3 Note : information de mesure reliant les entités utilisateur et item

La note est une information de mesure de la pertinence des résultats obtenus après un processus de recommandation. Une note est obtenue de deux manières (Harper et al., 2005) :

— par des algorithmes de prédiction (H. Li et al., 2015), qui prédisent les notes qu'un utilisateur pourrait attribuer à un item, pas encore noté. Cet item est contenu dans une liste dont les premiers sont les plus pertinents à recommander (Blandin et al., 2019). Soit D, l'ensemble des descripteurs associés à l'ensemble des items I et D_i,, un sous-ensemble des descripteurs de D décrivant l'item i. Le poids d'un descripteur d_i dans un item i peut être représenté par la fonction de présence,

définie par l'expression suivante :

$$presence(d_i) = \begin{cases} 1 \ si \ d_i \in D_i \\ 0 \ si \ d_i \notin D_i \end{cases} \tag{1.2}$$

— à partir du jugement de l'item par l'utilisateur. C'est une notation qui survient après une première recommandation à l'utilisateur qui donnera son avis sur l'item. Ces notations provenant de l'utilisateur peuvent prendre une forme numérique, de « j'aime » ou encore de commentaires. Une deuxième notation est obtenue à partir du jugement de l'item. C'est une notation qui survient après une première recommandation à l'utilisateur qui donnera son avis sur l'item. Ces notations provenant de l'utilisateur peuvent prendre une forme numérique, de « j'aime » ou encore de commentaires.

Les entités utilisateur, item et la note jouent un rôle pour la prédiction et la recommandation dans tous les types de SR. Ces types de recommandations ont été classés de différentes manières par des travaux existants.

1.2.4 Classification des SR

Comme nous l'avons vu précédemment, au cours des dernières années un nombre important de travaux de recherche ont traité la problématique de la recommandation. Ces travaux se sont focalisés sur des méthodes traditionnelles ou sur des méthodes plus récentes. Ces dernières sont issues de plusieurs domaines comme les sciences de l'information et, récemment, l'apprentissage automatique.

Les méthodes de recommandation peuvent être classées de différentes manières. Parfois plusieurs termes sont utilisés pour désigner une même méthode. Notre recherche se base sur les classifications les plus connues regroupant les principales méthodes de recommandation basées sur le FColl et le FCont (Burke, 2007 ; Rao, 2008).

Burke (2007) a ajouté trois autres méthodes : la recommandation basée sur la connaissance, la recommandation basée sur l'utilité et la recommandation basée sur la démographie.

Burke, souligne que ces méthodes, citées précédemment, sont des méthodes traditionnelles (FColl et FCont) bien qu'elles représentent des cas particuliers. Nous présentons dans la suite de cette partie la recommandation hybride, ensuite dans la prochaine partie, celles basées sur le FCont et sur le FColl et enfin, les SR sensibles au contexte.
Ci-dessous, la classification classique et les autres classifications de types de SR, selon Rao (2008). La classification de Burke est très intéressante, car il considère deux types principaux de SR qui sont ceux basés sur le FColl et ceux basés sur le contenu. Ensuite, les autres types proposés dans la revue de littérature sont issus de ces deux principaux.

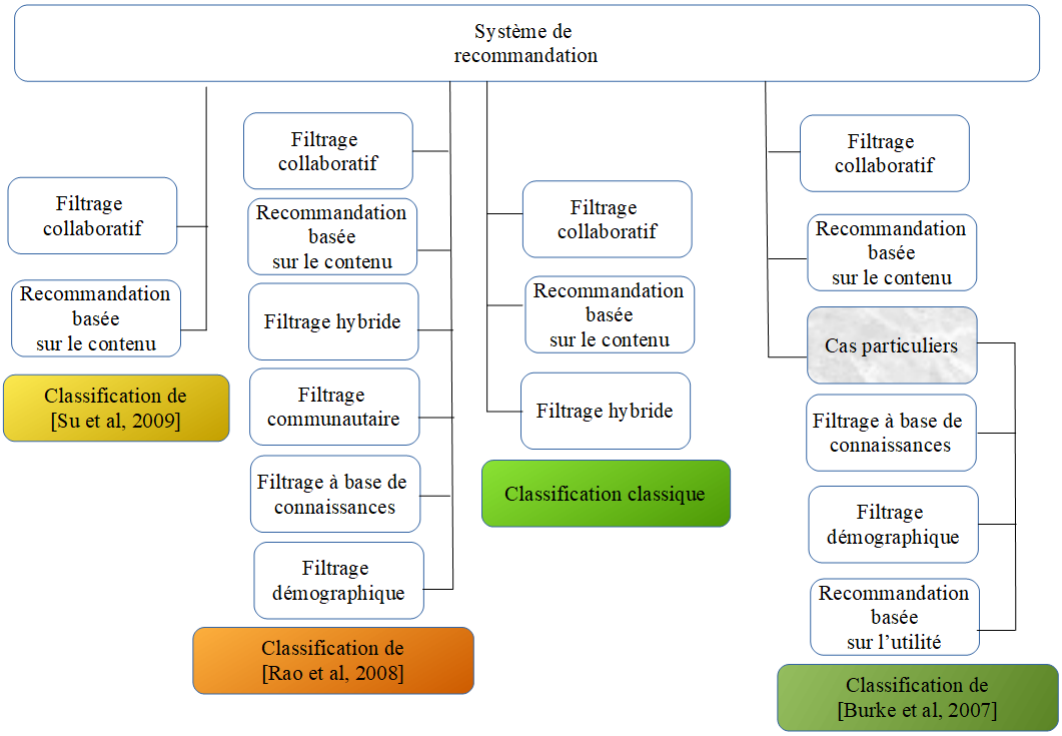

FIGURE 1.7 – Comparaison entre différentes classifications des systèmes de recommandations

Il existe plusieurs méthodes de filtrage pour la recommandation. Dans cette thèse, nous abordons celles basées sur le filtrage sur le contenu (FCont) et celles basées sur le filtrage collaboratif (FColl).

1.3 Recommandation basée sur le contenu

Les méthodes basées sur le filtrage du contenu (FCont) s'appuient sur les préférences de l'utilisateur et lui recommandent les items dont le contenu est similaire à ceux qu'il a aimés auparavant (Balabanović & Shoham, 1997 ; Adomavicius & Tuzhilin, 2005 ; Pazzani & Billsus, 2007).

La recommandation basée sur le Fcont peut être assimilée à un système de recherche d'informations exploitant le profil de l'utilisateur, car le Fcont parcours l'ensemble des informations concernant l'utilisateur en cherchant les préférences de ce dernier. Le profil utilisateur est composé de centres d'intérêts et sert à trouver des contenus présentant des métadonnées en parfaite adéquation. Cette technique est fondée sur la base d'analyse des similarités de contenu entre les différents items qui ont été précédemment consultés par les utilisateurs. Ce système utilise également les informations de retour d'expérience fournit par l'utilisateur (« feedback ») pour la mise à jour de son profil. Cela permet l'amélioration de la qualité des recommandations au cours du temps en incluant les retours d'expérience des utilisateurs dans le traitement de la recommandation. L'atout des systèmes de filtrage qui sont basés sur le FCont est la création d'une relation entre des items et un profil utilisateur. On peut citer en exemple le cas où le système utilise des techniques d'indexation et d'intelligence artificielle. L'utilisateur ne dépend pas des autres, ce qui lui donne la possibilité d'obtenir des recommandations même s'il est le seul utilisateur du système (Belloui, 2008). Dans le but de recommander par exemple des films à un utilisateur, le système procède à l'analyse des corrélations entre ces films et les films consultés dans le passé par cet utilisateur. Ces corrélations sont soumises à une évaluation en incluant les attributs comme le titre, le genre ou la durée.

Ces contenus sont décrits sur la base des méthodes traditionnelles ou de méthodes plus récentes. Nous allons présenter dans cette partie ces différentes méthodes et les technologies permettant leur conception.

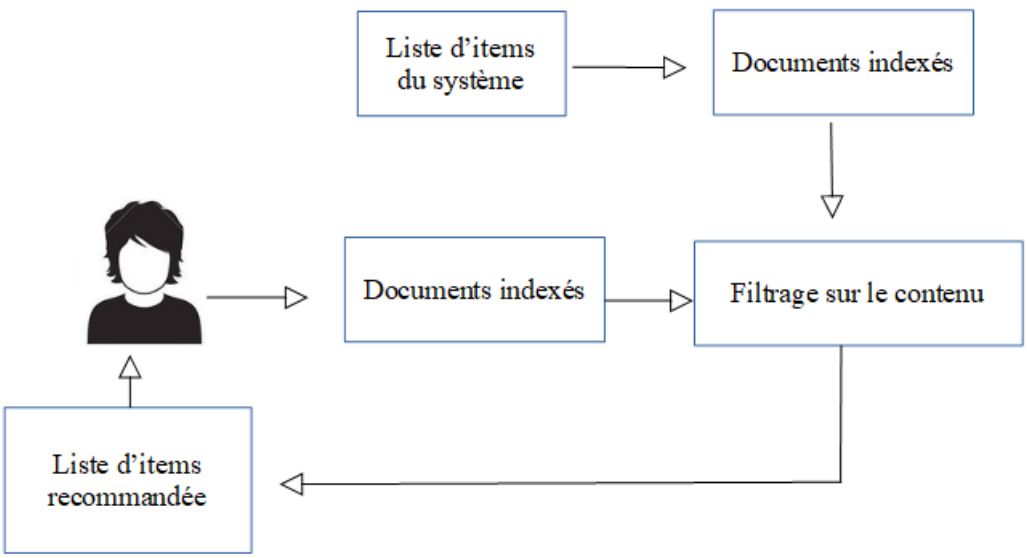

FIGURE 1.8 – Recommandation basée sur le contenu

1.3.1 Méthodes traditionnelles

Les méthodes traditionnelles permettent de faire une représentation des données et assurent le traitement de données sans se baser sur des modèles d'apprentissage automatique. Elles représentent en ensemble d'outils de gestion de données brutes, de connaissances ou encore de représentation sémantique.
Parmi ces méthodes traditionnelles de recommandation, nous allons aborder celles basées sur la connaissance, sur la sémantique ou encore sur l'utilité. Dans cette thèse, ces méthodes de représentation de données et de connaissances sont présentes dans notre contribution principale dans cette thèse.

Recommandation basée sur la connaissance

La connaissance est construite sur la base de l'information (nous allons voir dans le chapitre 2 le passage des informations aux connaissances). Tsuchiya (1995) présente la notion de connaissance comme suit : « l'information ne devient connaissance que lorsqu'elle est comprise par le schéma d'interprétation du receveur qui lui donne un sens ». Selon Penalva et Montmain (2002) la connaissance, à l'inverse de l'information, repose sur des systèmes de valeurs et de souhaits. Le traitement des SR correspond à l'idée

d'extraire des données disponibles sur le web, des connaissances utiles au sens de la recommandation et des préférences des utilisateurs.

La recommandation basée sur la connaissance utilise des connaissances bien précises, dont certaines caractéristiques d'items répondent aux préférences des utilisateurs. Ces systèmes à base de connaissances sont plus pertinents que d'autres méthodes de recommandation si les données disponibles sont limitées (dans le cas où le système ne peut pas compter sur l'existence d'un historique de l'utilisateur). Selon Piamrat et al. (2009), si un SR à base de connaissances n'est pas modélisé et conçu pour apprendre des notes ou des interactions de l'utilisateur, on observe différents raisonnements pour ce dernier :

1. un raisonnement à base des cas fondé sur la régularité du monde réel afin d'apporter des solutions aux problèmes en se basant sur des cas semblables rencontrés et résolus dans le passé. Piamrat et al. (2009) ont utilisé cette technique dans les SRs, ils mesurent les besoins ou les préférences de l'utilisateur qui correspondent aux recommandations possibles en se basant sur le comportement de consommation précédent ;
2. un raisonnement fondé sur des contraintes : corresponds à un type de système à base de connaissances. Cette recommandation à base de contraintes utilise les bases de connaissances prédéfinies qui contiennent des règles explicites sur la méthode consistant à associer les exigences des utilisateurs à des caractéristiques sur l'item. Par exemple, un utilisateur peut être intéressé par l'achat d'une voiture avec un ensemble de caractéristiques bien définies et dans un niveau de prix précis.

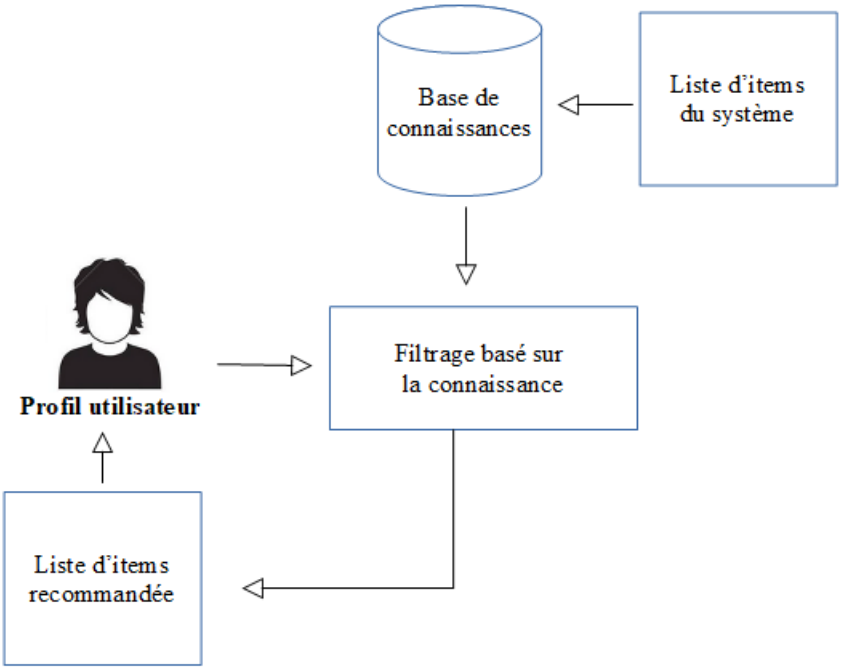

FIGURE 1.9 – Recommandation basée sur la connaissance

Recommandation basée sur la sémantique

Berners-Lee et al. (1998) présente en 1998 un article sur ce qui sera plus tard nommé le Web sémantique. Dans cet article, il parle du Web sémantique comme d'une extension du Web des documents, qui représentent une base de données mondiale, dont l'objectif est que toutes les machines puissent mieux lier les données du Web. La présentation du Web sémantique et de ses technologies sont abordées dans le chapitre 2 de ce manuscrit. La sémantique a été introduite avec plusieurs méthodes dans le processus de personnalisation pour des nouveaux sites Web de multilingues (Calabretto et al., 2009). La source externe de connaissances impliquées dans le processus de représentation est MultiWordNet (une base de données lexicale multilingue). Parmi les autres systèmes utilisant la sémantique pour la recommandation, on peut citer :

— SEWeP [7] est un système de personnalisation Web qui utilise à la fois les C-logs [8]

7. SEWeP est un système web de personnalisation qui intègre le processus de l'utilisation du web sémantique dans des sites web afin d'enrichir l'ensemble des recommandations fournies à l'utilisateur final. L'annotation sémantique de son contenu est réalisée à l'aide d'une hiérarchisation conceptuelle (taxonomie).

8. C-logs est une extension des logs dédiés au web qui en-capsulent la sémantique du contenu.

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 l'annotation sémantique des pages Web, dans le but d'avoir un vocabulaire uniforme et consistant ;

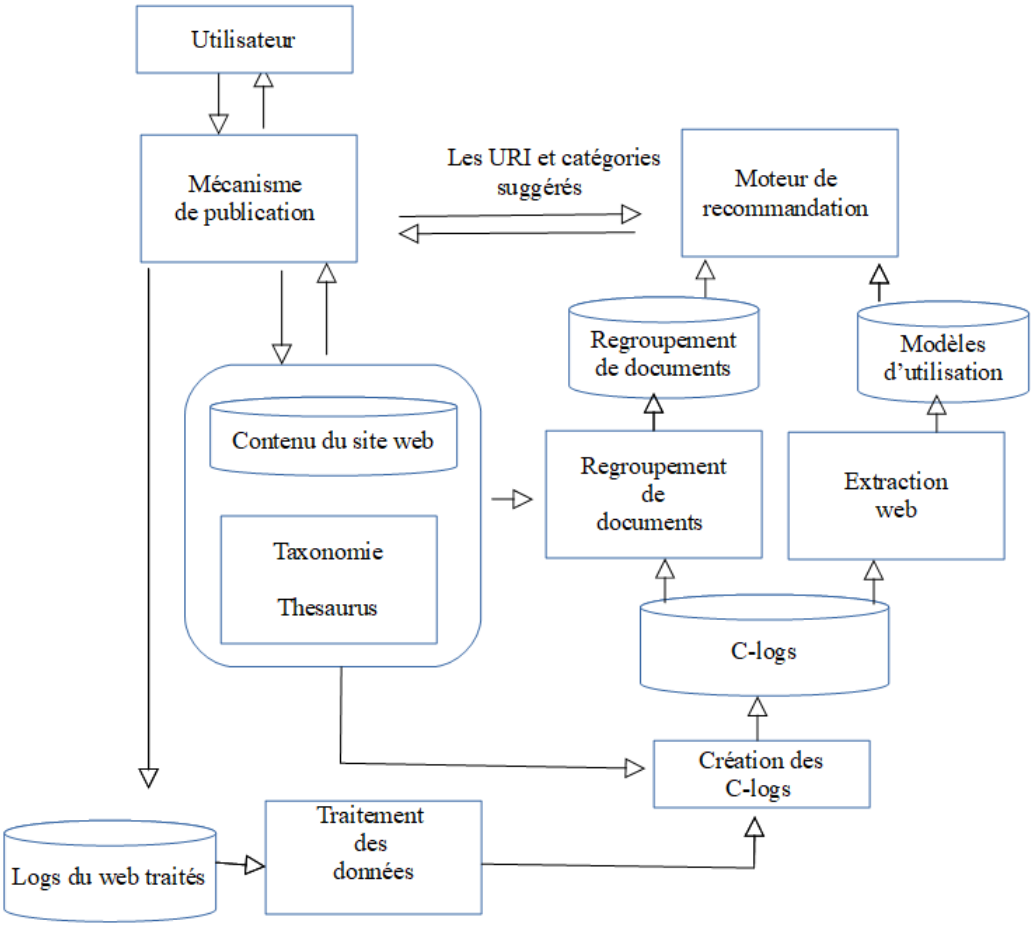

FIGURE 1.10 – Architecture système de SEWeP
(Eirinaki, Vazirgiannis, & Varlamis, 2003)

— Quickstep[9] est un SR d'articles de recherche académique (Middleton et al., 2002, 2004) et est basé sur l'ontologie dans le domaine de la recherche d'informations, Il permet d'indexer des articles. Son ontologie a été créée par des experts du domaine. Les concepts de son ontologie sont représentés comme des vecteurs d'exemple de produits.

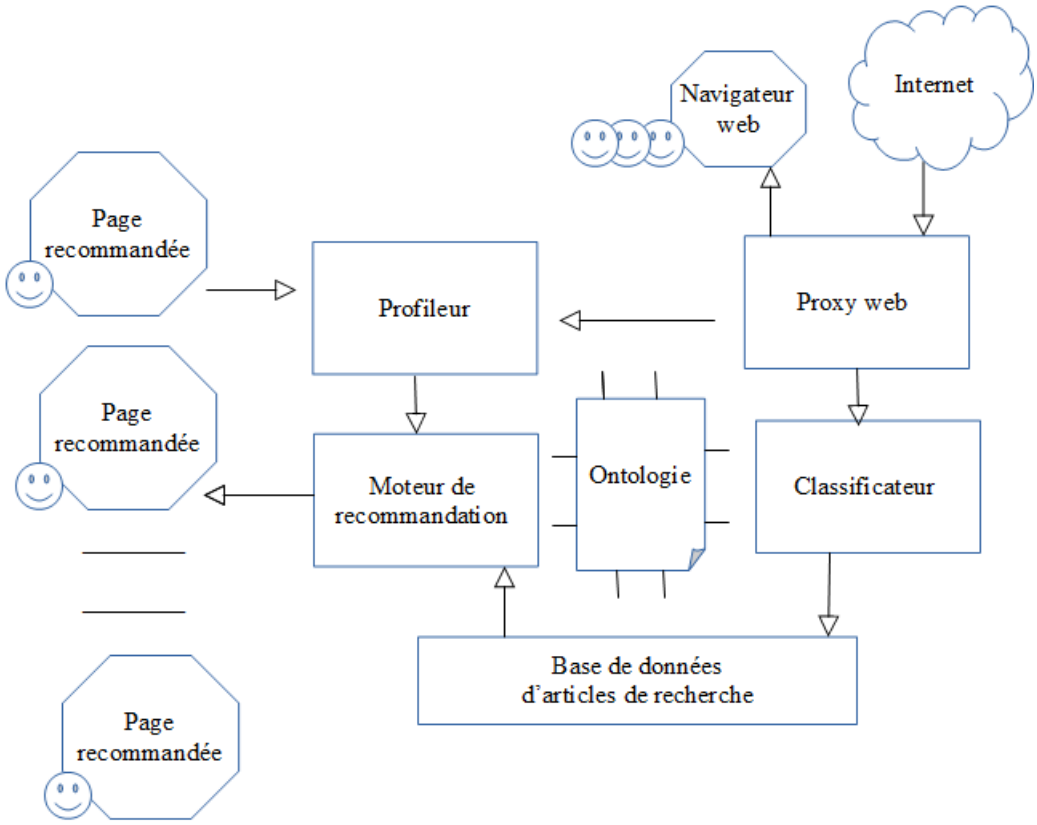

FIGURE 1.11 – Architecture système de QuickStep (Middleton et al., 2004)

9. Quickstep est un moteur de recommandation d'articles de recherche académique.

L'ontologie de Quickstep est basée sur la classification scientifique du projet DMOZ open directory (DMOZ open directory project). « Informed Recommender » (Aciar et al., 2007) utilise les avis des usagers sur les produits pour générer des recommandations. Le système procède à une conversion des opinions des utilisateurs dans une forme structurée. Cette conversion est effectuée en utilisant une ontologie de traduction, qui est exploitée pour la représentation et le partage de plusieurs connaissances. Ces méthodes ci-dessus ont donné des résultats pertinents et plus précis comparés aux méthodes traditionnelles basées sur le FCont.

La recommandation basée sur l'utilité

Cette méthode de recommandation est obtenue à partir de l'évaluation de l'utilité de chaque item pour l'utilisateur. Le grand problème de ce type de recommandation est la création d'une fonction d'utilité pour chaque utilisateur (Stolze & Rjaibi, 2001).
En effet, le profil de l'utilisateur est, dans ce cas, la fonction de l'utilité que le système va obtenir de l'utilisateur. Une manière de procéder est de demander aux utilisateurs de remplir un formulaire, ce qui est assimilable à un processus d'enquête de satisfaction sur un item donné. Ceci reste un processus très coûteux en termes de temps, mais efficace pour l'évaluation des items recommandés (Stolze & Rjaibi, 2001).

1.3.2 Méthodes récentes : recommandation basée sur l'apprentissage automatique

Avec le développement des nouvelles technologies de l'intelligence artificielle (IA), le processus de recommandation est entré dans une nouvelle aire (Lecun, 2016 ; Bengio et al., 2021). Cette évolution technologique a conduit à des méthodes d'hybridation entre ces nouvelles technologies de l'IA et les SR (Zhang et al., 2021).
Avant d'aborder la recommandation basée sur l'apprentissage automatique, nous allons présenter d'abord l'ensemble des techniques d'apprentissage automatique.
L'apprentissage automatique est de plus en plus utilisé dans les logiciels intelligents (Zoungrana, 2020 ; LeCun et al., 1998 ; Jadhav & Channe, 2016). Il s'agit d'un sous-domaine de l'intelligence artificielle qui a été étudié depuis la fin des années 1950 (Martens,

1959), il se concentre sur le développement de modèles (Figure 1.12) permettant de représenter certaines caractéristiques du monde qui nous entoure. Ils permettent aussi d'apprendre des propriétés statistiques de distributions des données traitées, dans le but d'accomplir une multitude de tâches (Ramkumar et al., 2018). Son lien avec l'intelligence découle de la capacité de ces modèles à extraire des informations pertinentes des données traitées lors d'un processus de mise à jour. Ce processus de mise à jour est appelé « training ». Il permet de réutiliser de manière thématique et efficace de nouvelles données jamais rencontrées auparavant. Selon Cunningham et al. (2008), un modèle est une fonction de décision qui, dans le cas de l'apprentissage supervisé, prend en entrée une valeur $x \in R$ *et* $(x_i, y_i)\ n_i{=}1$ où y_i est la valeur cible associée à x_i et qui renvoie une prédiction $f(x_i)$) moyenne de y_i. Lorsque la cible est discrète, il s'agit d'une tâche de classification.

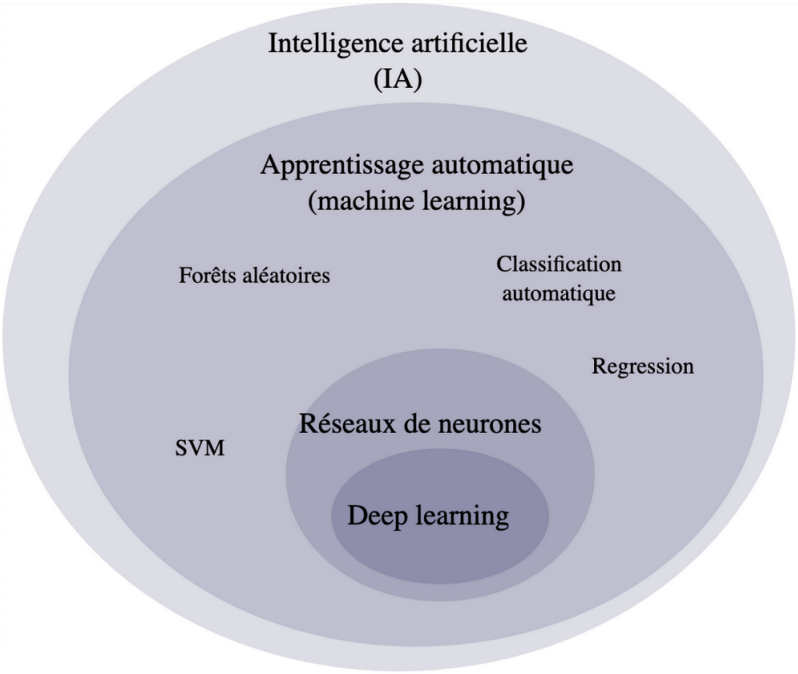

FIGURE 1.12 – Relation entre l'IA, l'apprentissage automatique et l'apprentissage profond

L'application potentielle des algorithmes de l'apprentissage automatique est vaste et le domaine semble très prometteur. Les modèles d'apprentissage peuvent être : supervisés, non supervisés ou encore par renforcement. Dans cette thèse nous allons-nous intéresser à l'apprentissage supervisé, l'apprentissage non supervisé et l'application de leurs algo-

rithmes respectifs dans un SR.

Apprentissage supervisé

Selon Marsland (2014), l'apprentissage supervisé est une démarche d'apprentissage permettant de faire des prédictions, en fonction d'un ou plusieurs modèles. D'après Marsland (2014), un algorithme d'apprentissage supervisé prend un ensemble connu de jeux de données en entrée. Ses réponses connues permettent de construire un modèle de régression ou de classification. En effet, un algorithme d'apprentissage entraîne un modèle dans l'objectif de générer des prédictions en guise de réponse à de nouvelles données ou à un ensemble de données de test (Geer, 2021). Il se base sur des algorithmes de classification et des techniques de régression pour développer des modèles prédictifs. En 2017, Denoyer et al. (2016) présentent un ensemble d'algorithmes se basant sur l'apprentissage automatique. Parmi ces algorithmes on a la régression linéaire, la régression logistique et les réseaux de neurones (Denoyer et al., 2016).

Pour Biernacki (1997), les modèles de classification permettent de prédire des réponses discrètes. Il souligne dans ses recherches qu'il est recommandé de faire une classification si les données peuvent être classées, étiquetées ou séparées en groupes ou classes spécifiques. Les applications les plus courantes ou les plus importantes de la classification comprennent l'évaluation du crédit bancaire, l'imagerie médicale et la reconnaissance de la parole. Les travaux de Crettez et Lorette (1998) montrent que la reconnaissance de l'écriture manuscrite utilise la classification pour identifier les lettres et les chiffres. Cette méthode de reconnaissance est utilisée pour vérifier si un courrier électronique est authentique ou non-spam, ou même pour détecter si une tumeur est cancérigène (Gherabi, 2018).

Les méthodes de classification abordées dans cette thèse sont celles qui sont orientées sur la prévision de recommandations pertinentes en analysant les items à proposer et en classant en fonction des tendances d'amélioration des ventes. Cette méthode est utilisée par exemple pour la classification des transactions frauduleuses des cartes bancaires (Géron, 2019). Selon Denoyer et al. (2016), il existe plusieurs méthodes de classification, parmi lesquelles : les arbres de décision, la régression logistique, les réseaux de neurones, les machines à vecteurs d'appui, l'analyse discriminante linéaire et enfin les K-plus proches

voisins (Ali et al., 2020 ; Guo et al., 2003). Dans ce travail de recherche, nous allons-nous intéresser à la technique des plus proches voisins.

Selon Abu-Nimeh et al. (2007), contrairement aux méthodes de classification, les techniques de régression prédisent des valeurs continues. Ils soulignent que la technique de régression linéaire est l'une des premières techniques d'apprentissage, elle est encore largement utilisée. Une régression linéaire permet de faire une modélisation de la relation entre deux variables en ajustant une équation linéaire des données observées (Denoyer et al., 2016). Par exemple, si des données sont collectées sur le degré de satisfaction des personnes après avoir visionné un ensemble de films. Dans cet ensemble de données, les films et les personnes satisfaites sont des variables. Par analyse de régression, on peut les relier et commencer à faire des prédictions en vue d'une recommandation.

Dans le domaine du traitement du langage, les travaux de Caelen et Villaseñor (1997) montrent que, l'entrée peut contenir un texte annoté fourni par des humains. Ce texte annoté est une métadonnée qui est fournie avec le jeu de données à la machine. En effet, les annotations peuvent être des balises de partie de la parole (balise PoS), des phrases et des structures de dépendance (Cleuziou et al., 2003). Par exemple pour déterminer si l'expression, « soutenir ma thèse » est une phrase nominale ou une phrase verbale, l'algorithme doit être formé en utilisant des phrases annotées telles que « soutenir ma thèse est un objectif » ou encore « soutenir sa thèse avant de se marier ». Dans le premier cas, l'annotation indique qu'il s'agit d'une phrase nominale et d'une phrase verbale dans le second cas.

Apprentissage non supervisé

Selon Fisher et al. (2014), en apprentissage non supervisé, la base de données comporte une collection de données non annotées sous la forme Xi avec i est dans [1, n] et x, le vecteur fonctionnel. Ils soulignent que l'objectif d'un algorithme d'apprentissage non supervisé est de créer un modèle de vecteur X en entrée et de faire une transformation des autres vecteurs sur la base d'un modèle.

Pour Halgamuge et Wang (2005), l'apprentissage non supervisé ou encore de classification automatique (« clustering ») est une technique très significative dans l'analyse de données. Elle permet d'identifier les groupes d'objets similaires dans un ensemble de

données sans pour autant connaitre la structure de ces dernières. Cette classification automatique est différente de celle de l'apprentissage supervisé, qui est plus orienté vers un concept de classement, permettant de déterminer les règles permettant de faire la séparation des objets dans un ensemble de données.
La classification automatique permet de procéder à la formation de groupes d'objets similaires. Cette notion de similarité est très importante dans la classification automatique, car permettant de regrouper des objets sous la base de critères bien définis. Soit un ensemble de points de vente : sur la base d'un ensemble de critères, toutes les partitions de cet ensemble sont acceptables comme classification (par exemple : groupe de points de vente de moins de 10 employés de l'ensemble). Il est très courant de définir le concept de similarité en s'appuyant sur la notion de dis-similarité. On considère que deux objets x et y sont similaires plus qu'ils soient proches au sens d'une mesure de dis-similarité. Ci-dessous, nous montrons la mesure de dis-similarité :

$$\text{Soit : un ensemble d'objets m : E x E/ (x, y)} \in E \times E \Rightarrow m(x, y) = m(y, x)$$
$$\text{avec m(x, y)} = 0 \Leftrightarrow x = y.$$

La définition de la mesure de similarité est très importante pour la classification de l'ensemble des objets dans le sens où l'apprentissage est non supervisé.
La logique de ces algorithmes d'apprentissage automatique correspond parfaitement au processus de recommandation. Leur utilisation dans une démarche de FColl peut être très intéressante.

Méthodes hybrides basées sur l'apprentissage automatique

Au fur et à mesure de l'évolution du domaine des SR, les chercheurs ont étudié l'utilisation d'algorithmes issus de l'apprentissage automatique. Il existe maintenant plusieurs algorithmes tels que k-nearest neighbor (Cherif, 2018), clustering ; le réseau Bayes (Friedman et al., 1997), pour n'en citer que quelques-uns. Certains types, utilisés dans des applications vont de la reconnaissance de formes dans les images (Navarro et al., 2019) aux véhicules autonomes (Vellinga, 2017). L'application potentielle des algorithmes d'apprentissage automatique est vaste et le domaine semble prometteur. Aujourd'hui, parmi

les plus grandes entreprises qui utilisent des modèles hybrides basés sur l'apprentissage automatique, on peut citer : Facebook (Tadlaoui et al., 2015) Google (J. Liu et al., 2010), ou encore Netflix (Steck, 2013).

L'apprentissage automatique est au cœur de nombreux produits et services essentiels de Facebook (Hazelwood et al., 2018). La grande quantité de données que Facebook possède sur ses utilisateurs en fait une source précieuse d'informations pour les SR. Tsang et al. (2020) ont proposé une méthode permettant d'interpréter et d'augmenter les prédictions des SRs. Ils proposent d'interpréter les interactions des utilisateurs du système à partir d'un modèle de recommandation source et de coder explicitement ces interactions dans un modèle de recommandation cible. Ils se sont basés sur une utilisation importante de la recommandation par apprentissage automatique : la prédiction « add-click ». Ils ont fait des interprétations sur les interactions qui sont à la fois informatives et prédictives, c'est-à-dire qu'elles surpassent de manière significative les méthodes de recommandation existantes, car exploitant l'ensemble de interactions de ses utilisateurs sur des algorithmes de prédiction appartenant à l'intelligence artificielle à la place des algorithmes des SRs traditionnels. De plus, cette même méthode de l'interprétation des interactions peut apporter de nouvelles idées dans des domaines allant au-delà de la recommandation, comme la classification de textes et d'images.

Dans leurs travaux, Tsang et al. (2020) ont identifié et exploité les interactions des utilisateurs qui représentent la façon dont un SR se comporte généralement. Ils proposent une nouvelle méthode, « Global Interaction Detection and Encoding for Recommendation (GLIDER) », qui détecte les interactions des utilisateurs qui s'étendent globalement sur plusieurs instances de données à partir d'une autre méthode de recommandation de référence considérée comme modèle. Puis ils encodent explicitement les interactions dans un autre modèle de recommandation cible. GLIDER y parvient en utilisant d'abord leurs travaux sur la détection des interactions neurales (NID) de Tsang et al. (2017). Avec leurs recherches sur la recommandation « add-click », ils ont constaté que les interprétations générées par GLIDER sont éclairantes et que les interactions globales détectées peuvent améliorer de manière significative les performances de prédiction du modèle cible. Comme notre méthode d'interprétation des interactions est très générale, nous montrons également que les interprétations sont instructives dans d'autres do-

maines : texte, image, graphique et modélisation.

Grâce à une méthode d'interception des données appelée LIME (Ribeiro et al., 2016) sur un lot d'échantillons de données, GLIDER encode explicitement les interactions globales collectées dans un modèle cible par le biais de croisements de caractéristiques éparses. Dans leurs expériences sur la recommandation « add-click », ils constatent que les interprétations générées par GLIDER sont éclairantes et que les interactions globales détectées peuvent améliorer de manière significative les performances de prédiction du modèle cible. Leur méthode d'interprétation des interactions étant très générale, ils montrent également que les interprétations sont instructives dans d'autres domaines : texte, image, graphique et modélisation. Ainsi, les méthodes qu'ils proposent viennent s'ajouter à celles proposées par Google.

En 2019, Google (Ie et al., 2019a) relève le défi de faire des recommandations basées sur des ardoises (tablettes électroniques) pour optimiser la valeur à long terme (VLT) [10] en utilisant l'apprentissage automatique (apprentissage par renforcement). La valeur à long terme est une mesure de croissance qui est mise à jour tout au long de l'apprentissage (Ie et al., 2019b). Ces travaux ont permis deux contributions. Une première contribution en développant SLATEQ, une décomposition de la différence temporelle basée sur la valeur et de l'apprentissage Q qui rend l'apprentissage par renforcement contrôlable avec des ardoises. Sous des hypothèses modérées sur le comportement de choix de l'utilisateur, ils montrent que la VLT d'une ardoise peut être décomposée en une fonction contrôlable de ses composantes VLT parallèlement.

En deuxième contribution, ils ont fait une démonstration de leurs méthodes en simulation et valident l'extensibilité de l'apprentissage par TD décomposé en utilisant SLATEQ dans des expériences en direct sur YouTube. D'après Piotte et Chabbert (2009), la méthode de recommandation de Netflix est basée sur le classement. C'est une méthode qui peut utiliser une grande variété de données pour arriver à un classement optimal de films pour chacun de ses utilisateurs. En effet, si on recherche une fonction de classement qui optimise la consommation, une base de référence évidente est la popularité des films. La raison en est claire : en moyenne, un utilisateur est plus susceptible de regarder ce que la plupart des autres utilisateurs regardent (Gomez-Uribe & Hunt, 2016). Cependant,

10. La valeur de long-terme est une prévision qui est prédite

la popularité est l'opposé de la personnalisation : elle produira la même commande de films pour chaque utilisateur. Selon Piotte et Chabbert (2009), l'objectif de Netflix est de trouver une fonction de classement personnalisée meilleure que l'indice de popularité des films, afin de satisfaire ses utilisateurs dont les goûts varient. Ainsi, Netflix propose un modèle d'apprentissage automatique qui sélectionne des exemples positifs et négatifs à partir de données historiques des utilisateurs et laisse un algorithme d'apprentissage automatique apprendre les poids qui optimisent leur objectif (Gomez-Uribe & Hunt, 2016). Selon Gomez-Uribe et Hunt (2016), le modèle d'apprentissage automatique proposé par Netflix correspond à leurs besoins. Néanmoins, Amatriain (2013) souligne un ensemble de problèmes lié à l'apprentissage automatique. Parmi ces problématiques on peut citer : « Learning to rank » qui est au cœur de scénarios d'application, tels que les moteurs de recherche ou le ciblage des publicités (Karatzoglou et al., 2013). Karatzoglou et al. (2013) notent que ce problème de « Learning to Rank » pourra être amélioré dans les années à venir. Ils soulignent cependant qu'une différence cruciale dans le cas des recommandations de classement est l'importance de la personnalisation : ils n'attendent pas une notion globale de pertinence, mais ils cherchent plutôt des moyens d'optimiser un modèle personnalisé.

Selon Piotte et Chabbert (2009), outre la popularité et les prévisions de classement, Netflix a essayé de nombreuses autres fonctionnalités. Certaines n'ont montré aucun effet positif, tandis que d'autres ont considérablement amélioré la précision de leur classement.

En effet, les chercheurs de Netflix précisent que de nombreuses méthodes d'apprentissage supervisées peuvent être utilisées pour le classement (Amatriain & Basilico, 2015). Amatriain et Basilico (2015) précisent que les choix typiques comprennent la régression logistique, les machines à vecteurs de soutien, les réseaux neuronaux ou les méthodes basées sur des arbres de décision comme les arbres de décision à gradient renforcé (GBDT). D'autre part, un grand nombre d'algorithmes spécifiquement conçus pour apprendre à classer sont apparus ces dernières années, comme RankSVM ou RankBoost.

Pour Piotte et Chabbert (2009), il n'y a pas de réponse facile pour choisir le modèle le plus performant dans un problème de classement donné. Ils soulignent que plus l'espace de variables est simple, plus le modèle peut être simple. Cependant, ils précisent qu'il

est facile de se laisser piéger dans une situation où une nouvelle variable n'a pas de valeur parce que le modèle ne peut pas l'apprendre. À l'inverse, de conclure qu'un modèle plus puissant n'est pas utile simplement parce que vous n'avez pas l'espace de variables exploitant ses avantages.

1.4 Recommandation basée sur le FColl

Le FColl est une technique basée sur le partage d'opinions entre les utilisateurs. Le terme a été introduit depuis moins de deux décennies et implémente le principe du « bouche à oreille » pratiqué depuis toujours par les humains pour se construire une opinion sur un produit ou un service (Schafer et al., 2007). Considérons, par exemple, que les amis de Jean trouvent que le dernier film qui est sorti en salle est un succès, il peut juger intéressant d'aller le voir. Si au contraire, la majorité de ses amis estiment que le film un échec, alors il se peut qu'il prenne la décision de ne pas y aller. Encore mieux, si Jean considère qu'il a toujours aimé les films recommandés par Valentin, que les films recommandés par Elodie l'ont toujours déçu et qu'Édouard recommande la totalité des films sans aucune distinction, au cours du temps, il décide alors de s'en tenir à l'opinion de Valentin. Ce type de recommandation se base sur des appréciations qui ont été données par un groupe d'utilisateurs sur un groupe de produits. Ces appréciations sont traduites en valeurs numériques et peuvent être des notes, des comptes d'achats effectués, des nombres de visites, etc. On peut citer deux grandes approches de FColl :

1. l'approche se référant aux utilisateurs (Resnick & Varian, 1997) consistant à faire une comparaison entre les utilisateurs et à retrouver ceux qui possèdent des goûts en commun, les notes d'un utilisateur étant ensuite prédites sur le base des informations de son voisinage ;
2. celle utilisant les produits ou articles (Sarwar et al., 2001) consistant à faire le rapprochement des différents articles appréciés par les mêmes personnes et de faire une prédiction des notes des utilisateurs en fonction des produits les plus proches de ceux qu'ils ont déjà notés.

Dans un système du FColl, il est nécessaire que les utilisateurs fournissent leurs évaluations sur les items qu'ils ont déjà utilisés, sous la forme des notes ce qui permet de construire

leurs profils et de les mettre à jour. Il n'y a pas d'analyse du sujet ou du contenu des items à recommander. Ce type de SR est très efficace dans le cas où le contenu des items est complexe, difficile à analyser, car l'utilisateur peut apercevoir divers domaines intéressants. En effet, le principe du FColl ne se focalise pas nécessairement sur la dimension thématique des profils, et n'est pas soumis à l'effet « entonnoir ». Dans les avantages du FColl, les jugements des utilisateurs n'intègrent pas seulement la dimension thématique, mais aussi d'autres caractéristiques relatives à la qualité des items tels que la diversité, la nouveauté, l'adéquation du public visé, etc. Le problème du FColl est que sa performance se base exclusivement sur la communication des évaluations (notes) données par utilisateurs. Dans le cas plusieurs items ont été utilisés et évalués par peu d'utilisateurs, ces items seront recommandés très rarement, même si ces utilisateurs leur ont donné des notes très élevées. Ce problème est connu sous le nom de problème de parcimonie (« sparsity problem »). De la même manière, lorsque les utilisateurs possèdent des goûts très différents en comparaison avec les autres, le SR n'a pas la possibilité de trouver des similarités entre utilisateurs et donc ne peut pas faire de bonnes recommandations.

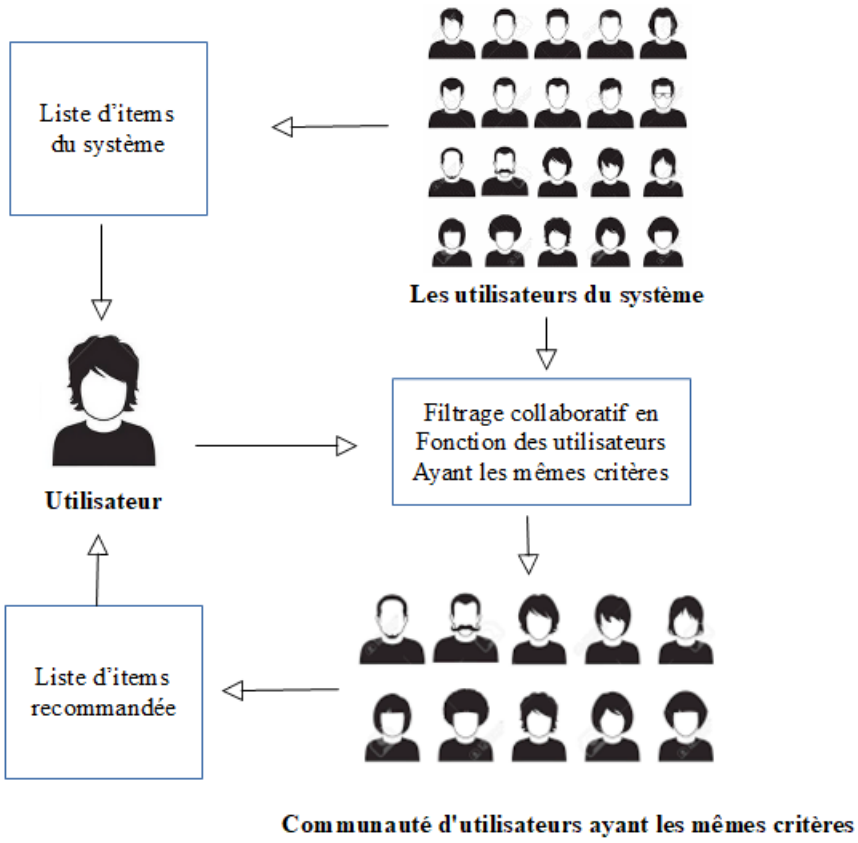

FIGURE 1.13 – Filtrage collaboratif

1.4.1 Les premiers systèmes de FColl

Le système Tapestry (D. Goldberg et al., 1992), cité au début de ce chapitre est le premier SR basé sur le FColl. Il se fonde sur l'opinion de ses utilisateurs pour faire un filtrage de messages électroniques. Tapestry stocke le contenu du message, son auteur et ses lecteurs et les avis des utilisateurs sous forme d'annotations (e.g. intéressant, bien). L'utilisateur peut définir son propre filtre en faisant la combinaison des mots-clés qui ont été associés au contenu avec des mots-clés générés à partir des annotations. Le terme collaboratif est alors introduit pour la première fois. Cela signifie que les utilisateurs collaborent entre eux pour s'entraider à filtrer leurs messages. Ce modèle de filtrage est connu sous le nom de « pull-active collaborative filtering » (Schafer et al., 2007), car c'est à la charge de l'utilisateur de rediriger vers lui les recommandations.

Après le développement du système Tapestry, les chercheurs se sont focalisés sur le potentiel que peut apporter le partage des opinions des utilisateurs dans un système de filtrage

d'information au sein d'une entreprise. Maltz et Ehrlich (1995) ont mis en place un « push-active collaborative filtering » qui donne la possibilité à un employé de « pousser » un document vers les personnes intéressées par son contenu. Ce type de recommandation s'est largement popularisé par la suite, particulièrement sur les réseaux sociaux où chaque utilisateur peut envoyer vers un groupe d'amis des documents susceptibles de les intéresser.

Dans ces exemples, on est face à un système de FColl actif qui demande, pour son bon fonctionnement, la connaissance mutuelle des utilisateurs pour partager leurs opinions. En effet, dans le système « pull active » l'utilisateur doit savoir à quelles opinions faire confiance, dans le système « active », l'utilisateur doit connaître les utilisateurs susceptibles d'apprécier l'item à partager. Dans les systèmes de filtrage collaboratifs, cette contrainte est supprimée et les groupes d'utilisateurs partageant les mêmes goûts sont formés automatiquement à partir d'une base de données contenant l'historique des préférences de tous les utilisateurs du système. Dans la liste des premiers systèmes de FColl automatisé, on peut citer GroupLenns (Resnick et al., 1994) utilisés dans les forums de discussions d'articles Use-net, Ringo (Shardanand & Maes, 1995) pour la recommandation d'albums de musique et d'artistes, et « Bellcore's Video Recommender » pour la recommandation de films.

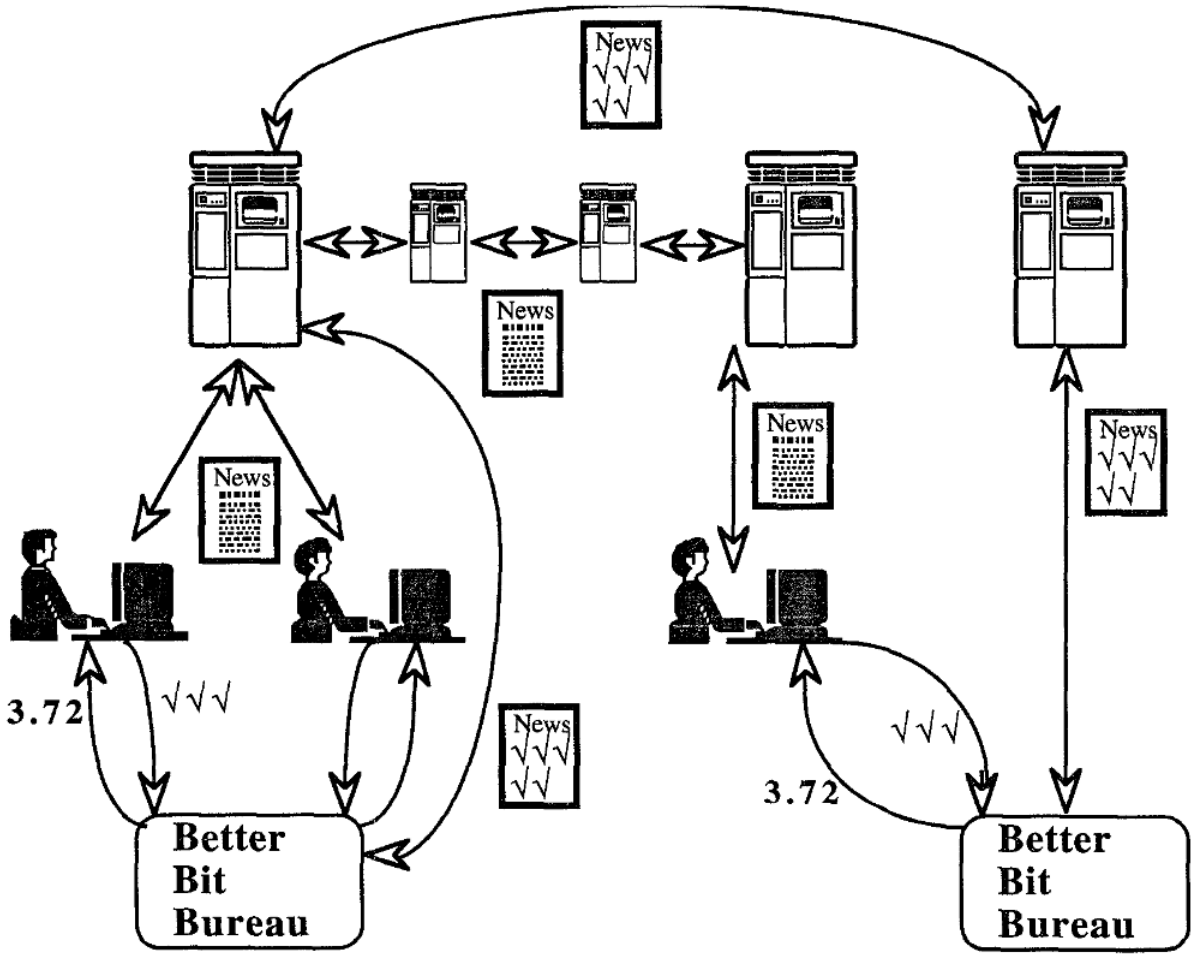

FIGURE 1.14 – Architecture GroupLens

L'architecture de GroupLens. Better Bit Bureau collecte les notations des clients, les communique via des serveurs d'articles et les utilise pour générer des prédictions de scores numériques envoyés aux clients. Les clients se connectent à un serveur d'informations local et peuvent se connecter à un « Better Bit Bureau » qui utilise le même serveur d'informations ou un serveur différent (Resnick et al., 1994).

1.4.2 Concept de similarité et ses métriques

Selon Xu (2007), le calcul de la similarité a pour objectif de déterminer si deux utilisateurs ou si deux items sont similaires. Ils présentent dans leur recherche plusieurs méthodes de calcul de similarité dont les deux les plus utilisées, car donnant les meilleurs résultats, sont : la méthode cosinus et la méthode du coefficient de corrélation de Pearson.

La méthode Cosinus est une mesure de similarité entre deux objets A et B de manière générale, très utilisée en recherche d'informations (Hamers et al., 1989). Cette mesure consiste à faire une représentation des deux objets sous forme de deux vecteurs $\overrightarrow{A}$ et $\overrightarrow{B}$ et de mesurer le cosinus de l'angle formé par ces deux vecteurs.'

$$Simiralite_{AB} = Cos(AB) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1.3}$$

Pour le FColl, chaque utilisateur u est représenté par un vecteur Au, ou $A_{ui} = r_{ui}$. Le cosinus de la similarité dans ce cas entre deux utilisateurs A et B, se fait sur l'ensemble des items notés par les deux utilisateurs.

$$r_{ui} = \frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \overline{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \overline{B})^2}} \tag{1.4}$$

Ce calcul peut être appliqué pour trouver le cosinus entre deux items. Le résultat du calcul du cosinus varie entre 0 et 1, si le résultat est égal à 1 cela signifie que les deux utilisateurs ou items sont similaires. Cependant, si le résultat est égal à 0, cela signifie que les deux utilisateurs ou les deux items n'ont rien en commun, ne sont pas similaires. La limite de cette mesure s'oriente dans le cas du FColl, car l'utilisation du cosinus ne tient pas compte de la variation dans le jugement des utilisateurs. La méthode du coefficient de corrélation de Pearson permet de calculer la corrélation statistique de Pearson entre deux vecteurs d'évaluation pour déterminer la similarité. Cette corrélation est mesurée sur la base des lignes de leur matrice d'évaluation respective. Ainsi, quand les items qui n'ont pas reçu d'évaluation sur les deux utilisateurs ne seront pas pris en compte, seulement ceux qui ont obtenu une évaluation seront pris en compte dans le processus de calcul. Le coefficient qui sera obtenu sera dans l'intervalle -1 et 1. Un coefficient qui est proche de -1 désigne une corrélation négative et celui qui tend vers 1, une corrélation positive. Si le coefficient tend vers 0, la similarité entre les vecteurs est inexistante, deux vecteurs non similaires. Soit X et Y, deux utilisateurs, la similarité entre ces derniers par la méthode du coefficient de corrélation de Pearson est exprimée par la formule suivant :

$$r_{ui} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{1.5}$$

1.4.3 FColl basé sur les voisins

Les SR basés sur le voisinage ont permis une automatisation du principe du bouche-à-oreille. Ces systèmes se basent sur l'avis de personnes qui partagent les mêmes idées ou d'autres sources pertinentes pour évaluer la valeur d'un item (film, livre, article, album, etc.), selon leurs propres préférences. Ainsi, dans le FColl basé sur le voisinage, les notes des utilisateurs stockées par le système sont utilisées pour prédire les notes de nouveaux items. Ce processus de prédiction peut être fait dans deux cas : recommandation basée sur le voisinage utilisateur ou recommandation basée sur les items.

Recommandation basée sur le voisinage utilisateur

GroupLens (Konstan et al., 1998) et Ringo (Shardanand & Maes, 1995), évaluent l'intérêt d'un utilisateur u pour un item i en utilisant les notes de cet item. Ces notes sont données par d'autres utilisateurs, appelés voisins, qui ont des habitudes de notations similaires. Les voisins d'un utilisateur A sont typiquement les utilisateurs B dont les notes sur les items sont plus proches de celles de A, sont les k utilisateurs B avec la plus grande mesure de similarité. Pour tout utilisateur B différent de A, les k plus proches voisins de A sont les k utilisateurs B avec la plus grande mesure de similarité par rapport à A. La note de l'utilisateur A sur l'item i peut être prédite par la moyenne des notes de B sur l'item i de ces voisins (Viola et al., 2019).
Un problème avec cette technique par la moyenne est qu'elle ne prend pas en compte le fait que les voisins peuvent avoir des niveaux différents de similarité. En effet, on peut prédire que la note de l'utilisateur A est plus à même de se rapprocher de la note de ses voisins avec une plus grande note de similarité à A. Cependant, si la somme des poids ne fait pas 1, les notes prédites peuvent être en dehors des valeurs autorisées. De ce fait, il est courant de normaliser ces poids de telle sorte que la note soit prédite. La technique par la moyenne ne considère pas le fait que les utilisateurs peuvent utiliser des notes différentes pour quantifier le même niveau d'appréciation sur un item. Ce problème est habituellement résolu en convertissant les notes moyennes des utilisateurs B sur l'item i en notes normalisées (Resnick et al., 1994). Ces méthodes de prédiction basées sur une moyenne de notes du voisinage résolvent essentiellement un problème de régression,

une orientation différente est la classification. Cela consiste à rechercher la note la plus probable que donnerait un utilisateur u à un item i, en prenant la valeur donnée le plus souvent par le plus proche voisin de u sur cet item et en considérant leur similarité avec A. Le vote sur l'item donné par les k plus proches voisins de A ayant donné une note appartenant à S peut être obtenu par la somme des valeurs de similarité des voisins qui ont donné cette note à l'item i. Une méthode de classification qui considère des notes normalisées peut aussi être définie en fonction d'un ensemble des valeurs normalisées possibles pour prédire une note.

FColl basé sur les items

Alors que les méthodes basées sur le voisinage utilisateur s'appuient sur l'avis d'utilisateurs partageant les mêmes idées pour prédire une note, les approches basées sur les items (Linden et al., 2003 ; Deshpande & Karypis, 2004) prédisent la note d'un utilisateur u pour un item i en se basant sur les notes de u pour des items similaires à i. Dans cette approche, deux items sont similaires si plusieurs utilisateurs du système les ont notés d'une manière similaire. Cette idée peut être formalisée comme suit. Soit I, l'ensemble des items notés par un utilisateur A qui sont similaires à un item i. La note prédite de A pour i peut être obtenue par la moyenne pondérée des notes données par A aux items de i.

1.5 Recommandation hybride

Selon Burke (2002), un SR hybride est un système qui combine deux ou plusieurs techniques de recommandations différentes, ces dernières peuvent être : la recommandation collaborative, la recommandation basée sur le contenu, la recommandation basée sur les données démographiques (Pazzani, 1999), ainsi que la recommandation basée sur la connaissance (O'Mahony & Smyth, 2007). Les SR démographiques ne permettent pas de faire des recommandations personnalisées, et la recommandation basée sur la connaissance suppose un échange d'informations sur les centres d'intérêt de l'utilisateur. Pour cette raison, nous nous intéressons dans cette thèse à l'hybridation entre la recommandation collaborative et la recommandation basée sur le contenu. Notre choix s'oriente vers

le système hybride parce que la recommandation basée sur le contenu et la recommandation collaborative ont souvent été considérées comme complémentaires (Adomavicius & Tuzhilin, 2005).

En effet, le filtrage sur le contenu permet de recommander les nouveaux items qui n'ont pas encore été évalués par aucun utilisateur. Tandis que le FColl ne peut recommander un item que s'il a été au préalable évalué par un certain nombre d'utilisateurs. Le filtrage sur le contenu nécessite la disposition de données sémantiques sur les items, en plus d'une étape d'analyse pour pouvoir les extraire et les représenter. Dans plusieurs domaines, le contenu sur les items est soit insuffisant (les livres sans résumés disponibles), soit difficile à extraire ou à représenter (musique, film). Le FColl ne requiert pas de contenu pour faire de la recommandation. La complexité d'un SR basé sur le contenu est liée à la difficulté avec laquelle sont extraites et analysées les données sémantiques sur les items. A titre d'exemple, si le système ne dispose que du genre des films comme information sur le contenu dans un SR de films, alors le modèle ne pourra intégrer que cette dimension. En plus, s'il s'avère difficile d'extraire automatiquement certaines propriétés sur les items, le filtrage sur le contenu est contraint de les ignorer dans ses algorithmes de recommandation. Par exemple, la qualité des données multimédia (image, vidéo ou audio) d'une page web peut représenter une information importante pour certains utilisateurs. Cette information est difficile à extraire automatiquement (Balabanović & Shoham, 1997 ; Schafer et al., 2007).

Le FColl permet l'évaluation d'une telle caractéristique puisqu'il se base sur les évaluations des utilisateurs. Par ailleurs, les recommandations produites par le filtrage sur le contenu pour un utilisateur donné souffrent d'un manque de diversité lié au problème de sur spécialisation. Le FColl est considéré par les chercheurs comme étant plus diversifié, produisant même des recommandations avec un effet de surprise, c'est à dire des recommandations pertinentes inattendues par l'utilisateur (Ye et al., 2019). L'hybridation de ces deux techniques, afin de traiter les insuffisances de chaque technique et profiter de leurs points forts, a fait l'objet de plusieurs travaux de recherche (Basu et al., 1998 ; Claypool et al., 1999). Avec la recommandation hybride, pour qu'un item soit recommandé à l'utilisateur deux critères doivent être satisfaits : son profil contenu doit être similaire au profil contenu de l'utilisateur, et l'item doit être apprécié par les voisins les plus proches

de l'utilisateur courant.

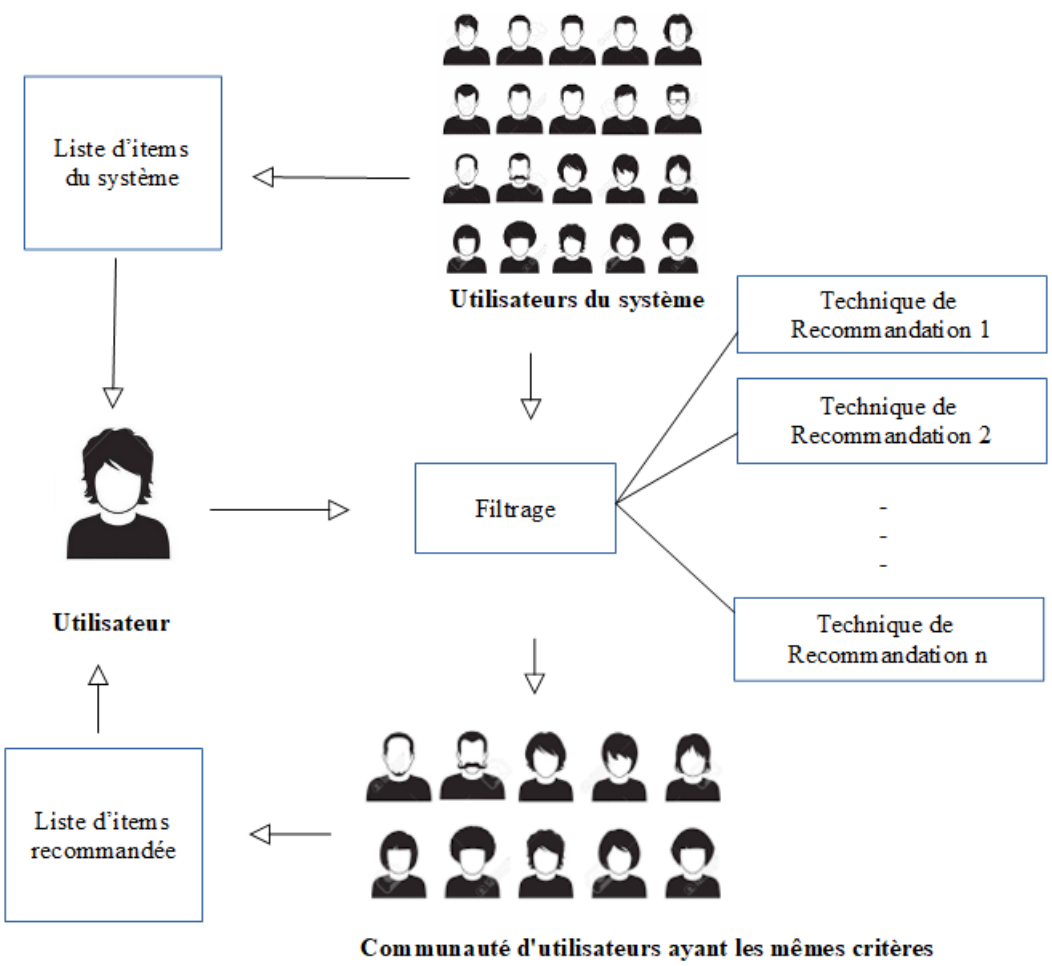

FIGURE 1.15 – Recommandation hybride

Il existe plusieurs procédés pour faire de l'hybridation et aucun consensus n'a été défini par la communauté des chercheurs. Toutefois, Burke (2002, 2007) a identifié sept méthodes différentes d'hybridation :

1. pondéré (« weighted) : le score ou le vote obtenu par chacune des deux techniques est combiné en un seul résultat ;
2. selection(« Switching ») : le système bascule entre les deux techniques de recommandation en fonction de la situation ;
3. mixte (« Mixed ») : les recommandations des deux techniques sont proposées simultanément ;
4. combinaison de propriétés (« Feature combination ») : les données issues des deux techniques sont combinées et transmises à un seul algorithme de recommandation ;
5. augmentation de propriétés (« Feature augmentation ») : le résultat d'une technique est utilisé comme entrée de l'autre technique ;
6. cascade : un système affine les recommandations données par l'autre système ;
7. « Meta-level » : une première technique construit un modèle qui sera utilisé comme entrée par la seconde technique.

Les systèmes hybrides sont définis par Burke (2002) comme étant une combinaison des méthodes traditionnelles précédemment présentées afin d'en pallier les limites, ces dernières sont actuellement les plus représentées dans la littérature, notamment, car elles sont jugées comme étant les plus efficaces (Schein et al., 2002). Selon Schein et al. (2002), un système hybride est organisé en deux phases : dans un premier temps, il faut effectuer de manière indépendante les filtrages des items utilisant des méthodes collaboratives ou par le contenu (ou autre). Dans un deuxième temps, il faut combiner ces ensembles de recommandations en exploitant des méthodes d'hybridations telles que des pondérations, commutations, cascades, etc.

Il existe plusieurs types de systèmes hybrides dans la littérature, ils reposent principalement sur deux approches présentées ci-dessus à savoir les méthodes basées sur le contenu et celles basées sur le FColl. La principale difficulté d'un système hybride consiste en l'hybridation elle-même. Comment, à partir des connaissances basées sur le contenu et

différents profils d'utilisateurs, pouvons-nous obtenir une recommandation efficace ? Plusieurs approches ont été proposées dont celle de Claypool et al. (1999) qui suggèrent d'hybrider les deux méthodes par l'utilisation d'une combinaison linéaire des deux mesures relatives à ces deux méthodes. Leurs travaux reposent sur ceux de Vogt et al. (1997) qui ont présenté une combinaison linéaire de scores retournés par différentes approches de recherche d'informations permettant d'améliorer de manière significative les résultats. Ainsi, dans le but de fournir à l'utilisateur un item, les auteurs calculent au préalable un score via l'approche collaborative et un score via l'approche basée sur le contenu.
La combinaison linéaire des deux notes va déterminer la note finale de l'item pour l'utilisateur. En outre Q. Li et Kim (2003) s'intéressent à la problématique d'hybridation en proposant une approche qui se base sur la construction de classes, en résolvant également le problème du démarrage à froid qui sera évoqué dans les inconvénients. Le principe de l'approche est décomposé en trois points :

1. l'utilisation d'un algorithme déposé aux utilisateurs « clustering » afin de grouper les items à proposer aux utilisateurs ;
2. calculer les distances entre les différents groupes précédemment constitués, mais également entre les items et les groupes, avec une mesure des cosinus améliorée et de Pearson ;
3. fournir une prédiction à un utilisateur en proposant des items proches de son voisinage.

Il faut noter que cette dernière approche ne propose pas tout à fait de méthodes qui s'appuient sur une approche collaborative ne regroupant pas directement des utilisateurs et ne construisant pas de profils. D'autres méthodes consistent à proposer deux techniques indépendantes de recommandation basées sur le contenu et à base de FColl. Cette hybridation permet de choisir les meilleures mesures de qualité selon Tran et Cohen (2000) qui proposent à un utilisateur une recommandation compatible avec ses précédentes évaluations.
Il existe un autre type de combinaison qui repose sur l'ajout de contenus dans les approches à base de FColl. La combinaison se fait au niveau du modèle collaboratif auquel on ajoute des ressources acquises via un système basé sur le contenu. On peut citer par

exemple les travaux de Soboroff et Nicholas (1999) qui proposent, la construction d'un modèle vectoriel contenant les profils utilisateurs se fondant sur le contenu des items. Quand la taille de la matrice est réduite par le biais de la méthode LSI « latent semantic indexing » (LSI) améliorant, les résultats obtenus avec un simple filtrage basé sur le contenu sont améliorés.

Une autre méthode qui a pour but de combiner les techniques fondées sur le contenu et les techniques collaboratives visant à produire un modèle de recommandation unique, c'est-à-dire sans avoir recours à la production de deux modèles devant par la suite être combinés. On peut citer les travaux de Basu et al. (2001) qui proposent une méthode à base de règles regroupant dans un même classificateur des notions basées sur le contenu et collaboratives. La partie collaborative est passive dans cette méthode car les auteurs récupèrent des informations des utilisateurs via le web sans interaction avec ces derniers. Popescul et al. (2001) ; Schein et al. (2002) proposent une approche semblable en utilisant un modèle probabiliste afin de combiner les techniques dans un modèle unique. Les auteurs proposent ici d'utiliser l'analyse latente probabiliste.

1.6 Autres méthodes de recommandation

En plus des FCont utilisant des méthodes traditionnelles et récentes, d'autres méthodes basées sur le FCont ont fait leurs preuves. Parmi ces dernières on peut citer les méthodes de recommandation utilisant des mots clés, des données pondérées, données démographiques, sensible aux contextes et celle basée sur des données communautaires.

1.6.1 Recommandation basée sur les mots-clés

La méthode de recommandation basée sur le FCont peut être appliquée à la recommandation de pages Web, de films, d'articles de presse, de restaurants, etc. Si nous prenons l'exemple d'un SR de document scientifiques basé sur le FCont, un utilisateur ayant l'habitude de consulter souvent des documents portant sur le domaine de l'informatique, le système lui recommandera des documents portant sur ce thématique.

En effet, ces documents disposent de mots-clés communs tels que : programmation, système informatique ou encore logiciel informatique. Ces mots-clés sont généralement

soit extraits sur la base d'une indexation automatique, soit attribués manuellement. Pour ce qui est des systèmes de recommandation de films ou de restaurants, le contenu est plutôt structuré et représenté par des métadonnées définies au préalable et valables pour tous les items (Pazzani & Billsus, 2007).

La représentation par mots-clés à la fois pour les items et pour les profils peut donner des résultats précis. La plupart des systèmes basés sur le FCont sont modélisés et 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.

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-clés chercheront seulement des documents dans lesquels les mots « français » et « impressionnisme » apparaissent.

1.6.2 Recommandation communautaire

La recommandation communautaire est utilisée dans les réseaux sociaux (Facebook, Twitter, etc). Le principe est le suivant : si des utilisateurs ont partagé des mêmes intérêts dans le passé, il y a de fortes chances qu'ils partagent aussi les mêmes goûts dans le futur. Ainsi, le moteur propose des recommandations à partir des relations de l'utilisateur dans le réseau social, et parfois en fonction de la confiance accordée par l'utilisateur à chacun de ses amis. L'exemple le plus populaire par rapport à cette recommandation est la notion des pages et des groupes que l'on retrouve dans une page Facebook. Le bouton « j'aime » de Facebook a une importance décisionnelle et a donné un succès grandissant des utilisateurs qui sont influencés par leurs amis (Garnine, 2020 ; Beam et al., 2018).

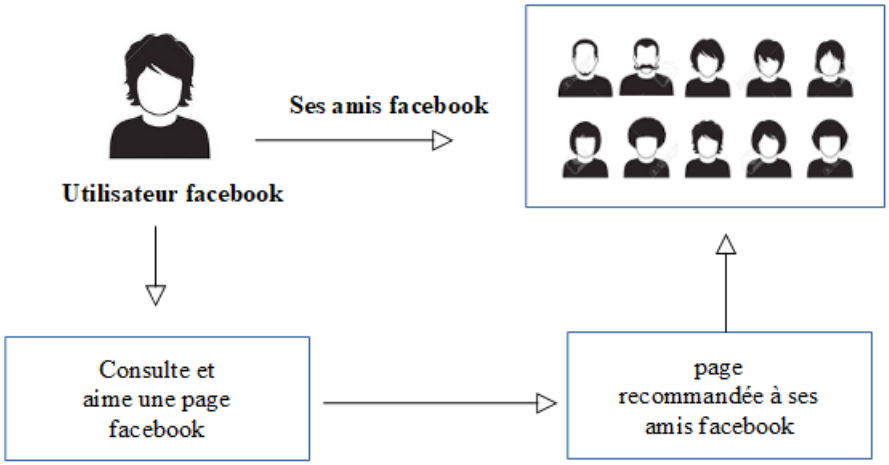

FIGURE 1.16 – Recommandation communautaire

1.6.3 Recommandation pondéré

Selon Montaner et al. (2003), les réseaux sémantiques permettent de sauvegarder la signification des mots permettant l'apprentissage du profil de l'utilisateur. Lops et al. (2011) soulignent que la principale motivation de telles méthodes est de procurer des SR intégrant l'aspect culturel et linguistique permettant d'interpréter le langage naturel et de raisonner sur son contenu.

Le système SiteIF de Stefani et Strapparava (1999) est un SR d'articles d'actualités multilingues intégrant des connaissances linguistiques pour faire des recommandations. Le profil utilisateur est représenté par un réseau sémantique où chaque nœud représente un mot lu par l'utilisateur, l'arc entre deux nœuds modélise la relation entre deux mots, un poids est associé à chaque nœud et à chaque arc représentant les différents niveaux d'intérêt de l'utilisateur.

Quickstep permet de recommander des articles de recherche (Middleton et al., 2004). Le système Quickstep utilise une ontologie décrivant la catégorie des articles de recherches pour modéliser les items. Une présentation plus détaillée de l'utilisation des ontologies (De Gemmis et al., 2015 ; Lops et al., 2011).

1.6.4 Recommandation sensible au contextes

Depuis les années 1990, internet a métamorphosé la manière de consommer et de vendre. Le commerce électronique est devenu une technique de commercialisation incontournable pour les entreprises de vente. La notion de contexte désigne tous les éléments pouvant se répercuter sur la compréhension d'une situation. Cette présentation contextuelle est utilisée pour mieux comprendre l'environnement.
En informatique, le contexte est un ensemble d'informations qui concerne un fait en liaison avec sa localisation conduisant au amenant le système informatique à s'adapter en conséquence et de à fonctionner. Selon Schilit et Theimer (1994), le contexte se base sur la localisation, sur les informations des personnes et les objets à proximité ainsi que les modifications susceptibles d'intervenir sur ces objets. Il a utilisé sa définition et proposé une méthode permettant de déterminer le contexte en répondant à trois questions : « Où êtes-vous ? », « Avec qui êtes-vous ? » et « Quelles sont les ressources à proximité de vous ? » (Schilit & Theimer, 1994). Quelques années plus tard, d'autres facteurs, comme la température, la saison ou encore l'heure ont été ajoutés à la liste de Schilit par P. J. Brown et al. (1997).

1.6.5 Recommandation démographique

La recommandation démographique est un système qui propose des items par rapport au profil démographique d'utilisateur (Safoury & Salah, 2013). C'est une méthode de recommandation qui consiste à partager les utilisateurs en plusieurs groupes en fonction d'informations démographiques telles que l'âge, la profession, le pays, la langue, le genre, etc.
Pour Bouchindhomme et Rochlitz (1992), le principe de cette méthode de recommandation est que deux utilisateurs ayant évolué dans un environnement similaire partagent des goûts communs, tandis que deux utilisateurs ayant évolué dans des environnements différents ne partagent donc pas les mêmes codes.
Plusieurs sites web utilisent la recommandation démographique pour proposer une offre de contenu personnalisé (J. Gupta & Gadge, 2015). Par exemple, les utilisateurs sont redirigés vers une plateforme Web en fonction de leur langue ou de leur pays. Ces approches

ont été très populaires dans la littérature du marketing, mais ont reçu peu d'attention dans le domaine des algorithmes de recommandation.

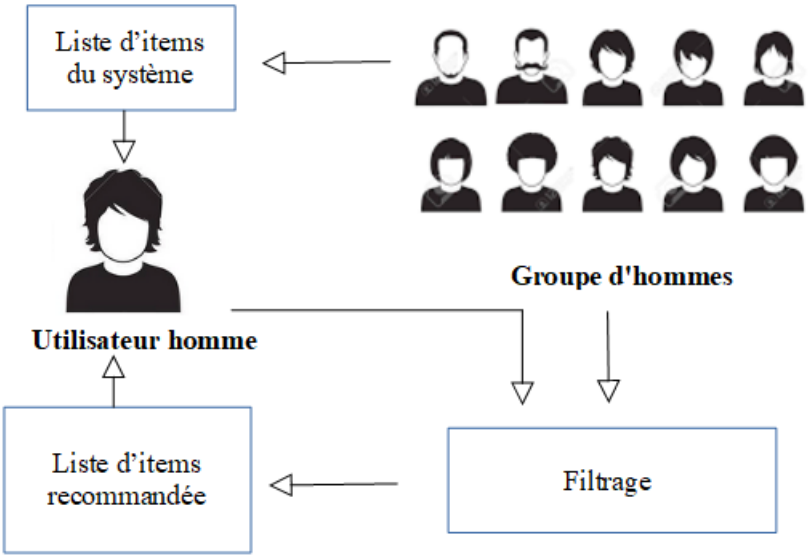

FIGURE 1.17 – Filtrage démographique

1.7 Avantages et inconvénients d'un SR

Les méthodes de recommandation traditionnelles présentent plusieurs avantages dont les plus importants sont :

1. la connaissance du domaine n'est pas nécessaire car le processus de recommandation se base uniquement sur les évaluations des items. En effet, seule la connaissance de l'utilisateur est requise (Soualah-Alila et al., 2014). Le caractère dynamique de ces systèmes est également un avantage car plus l'utilisateur va utiliser le système et plus la pertinence des items qui lui seront proposés sera fine ;
2. « Feed-back » implicite suffisant : habituellement un SR peut utiliser les écoutes sans avoir besoin d'une évaluation explicite(Sohail et al., 2014) ;
3. Adaptabilité : au fur et à mesure que la base de données des évaluations augmente, la recommandation devient plus précise. Un utilisateur peut se voir recommander des items de genres différents (Ezz & Elshenawy, 2020). Le moteur est fondé sur

le comportement de différents clients. Celui-ci analyse la navigation internet ou les achats des clients par exemple, en conclut des corrélations entre produits et présente les produits (ou solutions) les plus corrélés à votre recherche (souhait) ;

4. Cross-genre niches : le FColl se différencie par sa capacité à recommander aux utilisateurs ce qui se trouve en dehors du familier. C'est ce que Burke appelle : cross-genre niches(Burke, 2007) ;
5. les items sont plus simple à corréler que les utilisateurs et la recommandation est plus significative ;
6. la possibilité de recommander aux utilisateurs ayant un goût unique ou rare ;
7. trouver une corrélation entre un nombre limité d'items est mieux que de trouver une corrélation entre un nombre très grand d'utilisateurs ;
8. on peut travailler toujours à partir d'une base de données constamment mise à jour.

Les méthodes de recommandation traditionnelles travaillent sur la base des données du client à notre place. Cependant, l'utilisation de ces méthodes peut entrainer plusieurs problèmes :

1. le problème de démarrage à froid pour un nouvel item est un problème concernant le FColl, et non pas le FCont. Dans le contexte d'un filtrage à base de contenus, il suffit d'introduire l'item dans le système pour que celui-ci soit analysé et rentré dans le processus de recommandation. Dans le cas du FColl, il doit y avoir suffisamment d'évaluations pour que celui-ci soit pris en compte dans le processus de recommandation (Berrichi & Djouaher, 2020) ;
2. le problème de démarrage à froid pour un nouvel utilisateur est commun au filtrage basé sur le contenu et au FColl, c'est à dire qu'un nouvel utilisateur qui n'a pas encore accumulé suffisamment d'évaluations ne peut pas avoir de recommandations pertinentes(Berrichi & Djouaher, 2020). En effet un nouvel utilisateur dans le système doit avoir consulté ou fourni des appréciations pour un certain nombre de ressources avant que le système ne puisse lui fournir des recommandations pertinentes. Pour remédier à ce problème, il est demandé un certain nombre d'informations à l'utilisateur au moment de son arrivée (en nombre limité pour ne

pas rendre le système trop contraignant). Pour faire face aux problématiques liées au démarrage à froid, Netflix demande aux nouveaux utilisateurs de sélectionner quelques titres qu'ils aiment. Ces titres seront utilisés pour lancer les premières recommandations des nouveaux utilisateurs(Souilah, 2019) ;

3. le problème de démarrage à froid pour un système débutant survient lors du lancement d'un nouveau service de recommandation. Le système ne possédant pas d'informations sur l'utilisateur et sur les items, génère un problème de FColl. En effet un FColl ne peut pas fonctionner sur une matrice vide. La solution pour remédier à ce problème est de trouver des informations descriptives des items afin d'organiser le catalogue et d'inciter les utilisateurs à le parcourir jusqu'à ce que la matrice soit bien remplie renseignée (Lian et al., 2017) ;
4. le « shilling » : c'est l'action malveillante qui consiste à influencer la recommandation en créant des faux profiles (Lam & Riedl, 2004) ;
5. le « gray sheep » : c'est quand les utilisateurs ont des goûts atypiques, ils n'ont pas beaucoup d'utilisateurs en tant que voisins. Cela conduit à des recommandations pauvres (Ghazanfar & Prügel-Bennett, 2014). Ce problème survient souvent dans le FColl ;
6. le problème de redondance thématique des propositions soumises à l'utilisateur : un utilisateur ne se verra jamais proposer d'items qui n'auront pas été jugés similaires à ceux qu'il apprécie. Si un utilisateur ne s'intéresse qu'aux articles parlant de sport, il ne se verra jamais proposer un article politique. Bien que cet exemple volontairement exagéré ne se rencontre que rarement, il a l'avantage d'être pédagogique et d'expliciter cette limitation. Une manière d'y remédier est, par exemple, d'utiliser des algorithmes de classification permettant d'obtenir des propositions pseudo-aléatoires, comme c'est le cas avec les algorithmes génétiques précédemment évoqués (Sheth & Maes, 1993). Ces approches posent également le problème dans le cas d'un nouvel arrivant. En effet, un utilisateur qui n'aura jamais utilisé le système ne pourra pas se voir proposer des recommandations pertinentes, le système manquera d'informations. Un certain nombre d'heuristiques peuvent cependant résoudre ce problème, par exemple, en ne proposant pas de

recommandations avant d'avoir recueilli assez d'informations (Fischer & Stevens, 1990) ;

7. la limite de la recommandation basée sur le contenu (Ticha, 2015 ; Carré et al., 2009) : est qu'elle nécessite l'acquisition d'un nombre suffisant d'attributs décrivant les items. C'est pourquoi elle est appropriée dans le cadre des ressources textuelles ou quand les descriptions textuelles des ressources ont été entrées manuellement. Dans le cadre d'une ressources textuelles, un des problèmes provient de la méthodologie adoptée pour la classification de texte utilisée. En effet, deux ressources peuvent être similaires du point de vue de leurs attributs, mais avoir une quantité ou une pertinence non comparable. Dans le cas de la proposition répétitive d'un item à un utilisateur, le système doit éviter de ne recommander que des ressources similaires à celles qu'un utilisateur a déjà appréciées, cela empêche de recommander d'autres ressources que ce même utilisateur pourrait apprécier. Pour résoudre ce problème, il est possible de faire une proposition aléatoire d'items parmi les recommandations ;
8. la recommandation basée du FColl : le problème principal du FColl est le manque de données (Sahraoui, 2017 ; Oufaida & Nouali, 2008 ; Pessiot et al., 2006). En effet, pour les notes explicites, le pourcentage moyen d'items pour lesquelles les utilisateurs ont fourni une appréciation est très bas. Il en est de même pour les approches basées sur le contenu, le FColl rencontre le problème du démarrage à froid : avant que le système puisse fournir des recommandations pertinentes à un utilisateur, il faut que ce dernier ait fourni, implicitement ou explicitement, des appréciations pour un nombre suffisant de ressources. Un problème supplémentaire, par rapport aux recommandations basées sur le contenu, est que ce problème s'applique également aux nouvelles ressources introduites dans le système. Les solutions à ces problèmes se trouvent dans les approches hybrides, présentées ci-dessus.

*

* *

Dans ce chapitre, nous avons présenté les origines des SRs, leurs domaines d'application, leurs principales entités, leurs typologies ainsi que leurs méthodes de conception ont été

abordées. Leurs avantages et inconvénients ont été présentés. Nous avons présenté un ensemble de méthodes que nous qualifions de traditionnelles et un autre ensemble de méthodes, plus récentes, basées sur les technologies de l'apprentissage automatique.
Nos recherches sont focalisées sur : la recommandation basée sur le FCont et celle basée sur le FColl. Ces deux méthodes de recommandation sont complémentaires par rapport aux caractéristiques qu'elles présentent. Ainsi notre contribution principale s'oriente vers une hybridation.
Dans notre sujet de thèse, le contexte de recommandation est très important, car le SR doit inclure le contexte de la recommandation. Pour ce faire, le sens des données qui sont recueillies pour faire des recommandations doit être « compris » par le SR. La démarche de l'expertise humaine doit être intégrée dans l'algorithme en charge du raisonnement pour effectuer des recommandations. Dans le chapitre suivant, nous aborderons la démarche de l'expertise humaine et sa modélisation par les technologies du web sémantique.

Chapitre 2

Expertise humaine et représentation des connaissances

Sommaire

« Soyez curieux : c'est en cherchant à comprendre,
en voulant apprendre que vous devenez un expert,
que vous trouvez de nouvelles solutions. »
Albert Einstein, Physicien théoricien (1879 - 1955)
« Un expert est un homme qui a cessé de penser.
Pourquoi penserait-il, puisqu'il est un expert ? »
Frank Lloyd Wright, Architecte (1867 - 1959)

Dans le chapitre précédent, nous avons vu qu'un SR est un outil qui guide l'humain dans ses choix. Il est assimilé par ses utilisateurs à un expert qui participe au processus de recommandation et à la prise décision. Cette assimilation peut être justifiée par le fait que les algorithmes de traitement d'un SR sont basés sur une modélisation de la logique humaine. En d'autres termes, c'est une traduction informatique de l'expertise humaine.
Ainsi, pour mieux comprendre la logique de fonctionnement d'un algorithme de recommandation, il est nécessaire dans un premier temps de comprendre la logique humaine. A partir de cette compréhension nous pouvons décrire et représenter les connaissances des experts en utilisant les technologies du web sémantique.
Dans un premier temps, nous allons présenter l'expertise humaine, notamment le concept d'expert, les méthodes et les modèles d'analyse utilisés pour l'expertise humaine. Ensuite dans un deuxième temps, nous aborderons la représentation des connaissances en présentant le Web sémantique, l'extraction des connaissances et les mesures de similarités sémantiques.

2.1 Expertise humaine

Selon Rabinowitz et Glaser (1985) ; Chi et al. (2014), l'expertise humaine est spécifique à un domaine précis, ce n'est pas seulement une capacité générale. Pour Chi et al. (2014), l'expertise humaine n'est pas transférable d'un domaine à un autre c'est-à-dire qu'on ne peut pas être expert en tout.
Les méthodes employées dans le domaine du marketing font appel à des experts que ce soit dans le contexte de recommandation, de la prise de décision ou encore de la compréhension d'une situation (Décaudin et al., 2009 ; Bisseret, 1995) .

Pour mieux comprendre l'expertise humaine, il est nécessaire de présenter le concept d'expert, mais aussi d'aborder les méthodes et les modèles utilisés par les experts dans leur analyse.

2.1.1 Concept d'expert

Vieira (2014) définit un expert comme une personne apte à juger de quelque chose, connaisseur, savant ou spécialiste, personne dont la profession consiste à évaluer la valeur de quelque chose. Le concept d'expert est étudié dans plusieurs disciplines, allant du marketing, à la psychologie, de la jusqu'aux sciences politiques et sociales.

Le concept d'expert recouvre des professionnels reconnus pour la qualité de leurs connaissances (Vieira, 2014 ; Décaudin et al., 2009 ; Kumar et al., 2016). Ces professionnels possèdent des capacités et des expériences précises qui les rendent plus légitimes que d'autres à prodiguer des conseils ou des recommandations pertinentes à ceux qui en ont besoin. C'est la rareté de leurs connaissances et expériences qui en font des experts. Un expert est considéré aussi comme une personne possédant une expérience très riche, dans un domaine spécifique, mais aussi dans les domaines connexes. On ne peut pas imaginer qu'un expert en SR ne connaît rien sur les systèmes d'information, ou encore sur les sites e-commerce, ces domaines sont liés.

En marketing, le concept d'expert est utilisé pour étudier le contexte de la recommandation et de la prise de décision (Décaudin et al., 2009). Le concept d'expert désigne à la fois des professionnels reconnus pour la qualité de leurs connaissances ainsi que des clients qui sont des consommateurs avec une expérience dans une catégorie de produits ou services (Vieira, 2014 ; Décaudin et al., 2009).

En marketing, J. S. Armstrong (1991) définit un expert comme une personne possédant des connaissances approfondies dans un domaine ou dans une catégorie de produit et une expérience acquise.

En psychologie cognitive, de nombreuses études sur les comportements des experts ont été effectuées dans le but de les comparer à ceux des personnes ne possédant pas de compétences, considérées comme des experts novices (Rabinowitz & Glaser, 1985). Ainsi, la psychologie de l'expertise, le concept expert novice est souvent utilisé par les cogniti-

vistes pour étudier un domaine d'expertise particulier, notamment pour déterminer les besoins des sujets experts novices pour devenir des experts (Rabinowitz & Glaser, 1985). La technique qui a été utilisée par les chercheurs est d'identifier un groupe d'expert et un groupe d'experts novices. Après la phase d'identification, ces deux groupes auront la même série de problèmes à résoudre et la comparaison sur leur performance sera faite.
Par exemple, des travaux en psychologie cognitive se sont employés à identifier comment, dans un domaine de connaissance spécifique, les novices atteignent le niveau des experts (Rabinowitz & Glaser, 1985). D'autres ont abordé l'automatisation de l'expertise humaine pour valoriser le processus de recrutement (Gijana, 2011).
Parallèlement, des travaux connexes ont été menés pour comparer les recommandations des experts par rapport à celles obtenues par l'application de modèles simples basés sur des données brutes. Pour D. M. Armstrong (1978), les études de diagnostic d'évaluation psychologique, de prédictions financières ou encore les prédictions des experts ne sont pas supérieures à celles des modèles actuariels [1]. Selon Larreche et Moinpour (1983), les performances prédictives des modèles statistiques doivent toujours être considérées avec un certain recul, le recours à un jugement humain étant souvent nécessaire. Dans ce sens, Blattberg et Hoch (1991) préconisent des prévisions basées sur les modèles statistiques combinées avec un jugement d'experts. Selon une étude sur l'utilisation double des experts et des modèles de base de données peuvent améliorer chacune des décisions prises isolément, car ces deux approches se complètent (Larreche & Moinpour, 1983).
Ainsi, dans l'étude effectuée dans le domaine des sciences politiques par Voss et Post (1988), les performances obtenues par des experts en chimie sont équivalentes à celle de novices en sciences politiques. Cependant, Cheng et Holyoak (1985) soulignent la force de raisonnement par analogie comme une possibilité de transfert de connaissances entre domaines, ce qui amène à croire que le transfert de méthodes, pour certains domaines est possible (Gick & Holyoak, 1987).

1. Relatif au calcul des opérations financières ou d'assurances

2.1.2 Méthodes d'analyse de l'expertise humaine

Le processus d'analyse de l'expertise humaine est basé sur une ou plusieurs méthodes. Nous allons aborder quelques-unes les plus utilisées par les experts pour l'analyse de données de manière générale. On peut ainsi en retenir deux, à savoir la méthode de prévision Delphi, qui est reprise dans plusieurs travaux en marketing (Vernette, 1994), et la méthode de la politique de jugement (Larreche & Moinpour, 1983 ; Chakravarti et al., 1981).

Méthode Delphi

La méthode Delphi est développée par la Rand corporation et théorisée à partir des années 60 (Dalkey, 1968b ; Linstone et al., 1975 ; Vernette, 1997, 1994) pour être exploitée dans plusieurs domaines : le marketing (Mohammadian Mahmoudi Tabar et al., 2021), la santé (Kastein et al., 1993), ainsi que l'enseignement (Fazio, 1985).
La méthode Delphi consiste à identifier un groupe d'experts sur un problème particulier. Chaque expert est interrogé individuellement pour obtenir un avis détaillé. Puis les avis sont comparés afin de constater les convergences et trouver un éventuel consensus. La réponse statistique du groupe d'experts à chaque question est communiquée aux membres du groupe dans le but d'initier une nouvelle interrogation individuelle et inciter chaque expert à réexaminer son jugement pour un consensus ; les itérations sont censées se renouveler jusqu'à l'obtention du consensus. La littérature (Jain, 1985) évalue à trois le nombre d'itérations généralement nécessaires à cette phase finale. Dans les expériences poursuivies par Dalkey (1968a) pour tester la valeur de cette méthode, la réponse de groupe après une discussion en face à face est dans la plupart des cas moins précise qu'une simple médiane des estimations individuelles sans discussion. D'autre part et sur un autre plan, la méthode Delphi a fait l'objet d'une variante, Delphi leader où les experts sont remplacés par des leaders d'opinion (Vernette, 1997).
Si la méthode Delphi n'est pas largement reprise en marketing, cela est essentiellement dû aux critiques qui lui sont adressées (Welty, 1971). En effet avec cette méthode, les experts possédant les capacités similaires (équivalentes) ont plus de chances de proposer des solutions différentes sur la même question. Si différents experts aboutissent à des

réponses fortement différentes pour la même question, cela aboutit à un consensus artificiel (Linstone et al., 1975).

D'après Sackman (1974), les résultats de cette méthode souffrent d'un manque de validité. Elle possède également trois principaux défauts. Primo, elle est lourde et fastidieuse tant pour les analystes que pour les experts. Deuxièmement, elle apparaît, plus intuitive que relationnelle. Troisièmement, seuls les experts qui sortent de la norme sont amenés à justifier leur position : Sackman (1974) recommande d'exploiter les différentes zones de divergences plutôt que de les ignorer. Pour U. G. Gupta et Clarke (1996), Delphi n'exploite pas la manière dont est mesurée l'expertise des évaluateurs, et le postulat de la supériorité de l'opinion d'un groupe par rapport à l'opinion individuelle est, de leur point de vue, discutable. Néanmoins, la technique Delphi a un avantage dans certains plans théoriques. En effet, dans les différents travaux qui ont fait la comparaison de plusieurs techniques, Delphi a souvent présenté les meilleurs résultats (Vernette, 1997).

En particulier, U. G. Gupta et Clarke (1996), soulignent les avantages de la méthode Delphi qui sont :

- l'anonymat permettant d'éviter l'influence d'un personnage fort (supérieur hiérarchique, personnalité imposante) ;
- pas de contrainte géographique ;
- temps de réflexion laissé aux experts pour répondre et faire mûrir leur réflexion ;
- peu coûteux ;
- bonne acceptabilité des résultats.

Dans cette thèse, nous nous sommes inspirés de cette méthode, car elle présente l'avantage de faciliter la proposition de recommandation de manière collaborative. Cette méthode s'adapte parfaitement avec celle utilisée pour le traitement des données issues des enquêtes mystères. Ainsi la méthodologie adoptée dans cette thèse face aux résultats issus des enquêtes mystères est la suivante :

- la définition de l'objet sur lequel portera la méthode. L'objet correspond à la problématique que vont examiner les experts et les questions liées à ce problème ;
- le choix des experts est effectué selon différents critères, selon leur niveau de connaissance de l'objet et leur indépendance ;
- le questionnaire sera élaboré avec des questions bien ciblées, précises et quanti-

fiables ;

— le questionnaire sera administré et les réponses traitées. Le premier questionnaire servant de base sera enrichi à chaque itération par des résultats et commentaires qui sont générés antérieurement.

Politique de jugement

Dans la littérature, cette politique de jugement est appelée : « policy capturing », « judgmental policy capturing » ou encore « bootstrapping du jugement » (J. S. Armstrong et al., 2001).C'est une méthode qui relève initialement de la psychologie cognitive. Elle a été appliquée dans de nombreux domaines, mais comme le soulignent Arregle et al. (2000), « rarement en gestion » : en finance (Slovic, 1972) dans le domaine des ressources humaines (Klass et al., 1991) et en stratégie (Hitt et al., 1995). Dans le domaine du marketing peu de recherches spécifiques à la méthode du « judgmental policy capturing » ont été développées (Batsell & Lodish, 1981 ; Arregle et al., 2000). Autrement, les aspects du marketing utilisant cette méthode proviennent de travaux issus d'autres disciplines et restent mineurs (Webster & Trevino, 1995).

Les recherches sur la recommandation en psychologie dans les années 1950 constituent les premières origines de la politique de jugement. Pour Hammond (1955), elle fait partie d'une technique de modélisation structurelle ayant pour principe : l'utilisation de modèles mathématiques pour la description des stratégies de jugements. Ces stratégies concernent les recommandations qui peuvent être modélisées par une équation mathématique qui relie les jugements d'un expert à la réalité à travers un certain nombre de variables. Les experts utilisant cette technique proviennent du milieu professionnel. Leur démarche méthodologique s'appuie sur des techniques de régression (Aiman-Smith et al., 2002) entre une série de jugements émis par des individus sur des scénarios et valeurs des variables composant ces scenarii (Arregle et al., 2000). Dans cette démarche d'évaluation, les individus éprouvent de grandes difficultés à relativiser et combiner les informations pertinentes pour leurs jugements (Slovic & Lichtenstein, 1971). Ainsi, la description que fournissent les personnes responsables de décisions de leurs politiques est très souvent inexacte (Aiman-Smith et al., 2002). A. K. Brown et al. (2005) précisent que dans toute activité, les individus attribuent une importance inégale des facteurs différents, ils ne

savent pas, de manière explicite, leur affecter un poids spécifique. Pour Brown et ses collègues, la méthode de modélisation structurelle propose une démarche de la politique de jugement basée sur la révélation de la théorie d'usage. Cette dernière est fondée par opposition, à la théorie professée à partir de l'observation de jugement des experts sur une série de cas : la démarche commence par l'étude des prévisions des experts puis « un saut en arrière » est effectué en vue d'inférer le raisonnement que les experts ont eu à faire pour déterminer leurs provisions (Green et al., 2007). Par conséquent, pour Zedeck (1977), la politique de jugement permet de révéler la politique qui sous-tend les jugements des experts dans le sens où elle permet d'évaluer la manière dont les décideurs utilisent les informations quand ils émettent un jugement. Elle révèle la démarche des décideurs qui « pèsent, combinent ou intègrent l'information ».

Ainsi, la problématique qui résulte de l'analyse du politique de jugement est la « politique de notation du jugement » pour chaque évaluateur (Zedeck, 1977). En effet, pour Zedeck (1977), la politique de jugement de l'expert sur le thème choisi est ainsi révélée ou capturée sous forme d'une équation d'où le nom de « policy capturing » (Westenberg & Koele, 1994). La compréhension des mécanismes psychologiques à l'origine de ces modélisations reste floue. D'autre part, les processus d'évaluation collectifs ne sont pas considérés.

En parallèle à ces deux méthodes, Larreche et Moinpour (1983)comparent les performances de différentes méthodes de recueil du jugement et développent un outil de mesure qui sert à sélectionner les experts de manière optimale. Les résultats obtenus montrent que, par ordre croissant, les approches qui permettent d'avoir un jugement de bonne qualité sont la moyenne des jugements individuels. Le jugement obtenu par consensus par la technique Delphi est celui identifiés sur la base de mesures externes.

Le tableau 1 est un résumé des deux méthodes d'analyse de l'expertise humaine présentées en amont.

Noms	Domaines	Apports	Limites	Références
Méthode Delphi	- Marketing - Recrutement - Système d'information - Management - Enseignement	- Anonymat permettant d'éviter l'influence d'un personnage fort (supérieur hiérarchique, personnalité imposante) - Pas de contrainte géographique - Temps de réflexion laissé aux experts pour répondre et faire mûrir leur réflexion. - Peu coûteux - Bonne acceptabilité des résultats	- Technique longue dans le temps (il faut que tout le monde réponde. . .) - Ne retient pas les idées extrêmes bien que parfois novatrices et originales. - Biais de sélection dans la constitution du panel d'experts. - Absence de débat entre les participants.	- Baillette, P., Fallery, B., et al., (2013) - Dalkey, et al., (1963) - Linstone, Turoff, et al., (1975) - Kastein, et al., (1993) - Fazio, 1985
Politique du jugement	- Psychologie cognitive - Système d'information - Marketing - Finance - Resources humaines - Finance - Mathématique	L'utilisation de modèles mathématiques pour la description des stratégies de jugements. permet de révéler la politique qui sous-tend les jugements des experts dans le sens où elle permet d'évaluer la manière dont les décideurs utilisent les informations quand ils émettent un jugement.	La compréhension des organismes psychologiques à l'origine de ces modélisations reste floue. les processus d'évaluation collectifs ne sont pas considérés.	- Larreche et Moinpour (1983) - Zedeck, 1977 - Arregle et al., (2000) - (Slovic, 1972) - Hitt, Tyler, et al., (1995) -(Batsell et al 1981) et (Arregle et al., 2000)

TABLE 2.1 – Comparaison entre les méthodes d'analyse de l'expertise humaine

Les méthodes citées précédemment sont accompagnées d'un ensemble de modèles de traitement.

2.1.3 Modèles d'analyse de l'expertise humaine

Afin d'analyser des données, l'expert peut se baser sur une succession de modèles d'expertise. Parmi ces modèles on peut citer : le modèle d'acquisition des données, le modèle d'agrégation des données selon le contexte et le modèle de traitement des données.

Modèle d'acquisition des données

Le modèle d'acquisition des données est un modèle établi par l'expert dans le but de capturer toutes les données qui lui seront nécessaires pour l'analyse et l'interprétation d'un cas d'étude. Ces données peuvent être capturées par divers modules d'acquisition. Ces modules peuvent être physiques ou logiques (logiciels), permettant de collecter des données (Cottet, 2020). Les modules physiques sont tous les dispositifs matériels qui sont capables de fournir des données de contexte, tel que le GPS [2], pour la détermination des coordonnées d'un utilisateur. Les modules logiques fournissent les informations contextuelles via des applications ou des services.

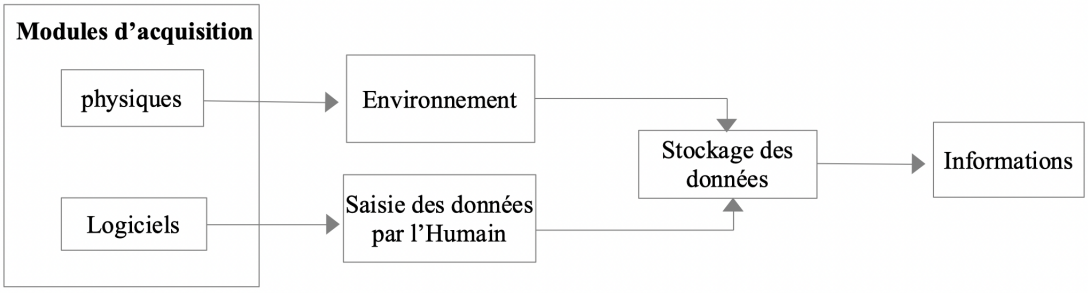

FIGURE 2.1 – Acquisition des données

Dans le cadre de la thèse, le modèle d'acquisition des données est défini à la suite de l'obtention des résultats d'enquêtes mystères. De fait, pour réaliser le modèle d'acquisition de données du SR que nous proposons dans cette thèse, nous établissons un modèle de sélection des données parmi celles acquises durant les enquêtes mystères.

2. Système de localisation par satellite

Modèle d'agrégation des données selon le contexte

Le modèle d'agrégation correspond dans la démarche d'analyse de l'expert à l'étape d'interprétation des informations contextuelles obtenues des modules d'acquisitions (Barkat, 2017). Cette étape a pour but l'analyse et la transformation des données brutes obtenues durant l'acquisition des données, de sources différentes, pour les convertir dans des formats plus faciles à manipuler et à exploiter par l'ordinateur (Barkat, 2017). En effet, si l'on prend l'exemple des coordonnées géographiques d'un point de vente décrits en fonction de la latitude et de la longitude (e.g. la latitude : 47.495030 et la longitude : 6.803110), on peut constater que ces coordonnées peuvent être transformées en adresses littérales (e.g. 4 Place Tharradin, 25200 Montbéliard) pour qu'elles puissent être exploitables par l'humain.

Après la capture et l'interprétation des informations contextuelles, ces dernières doivent être organisées pour faciliter leur exploitation (Luo & Seyedian, 2003). Ainsi, pour Luo et Seyedian (2003), la gestion du contexte englobe l'organisation et la représentation formelle des informations contextuelles en fonction du modèle. Le choix du modèle se base sur les mécanismes choisis pour l'adaptation du système selon le contexte dont les données ont été collectées.

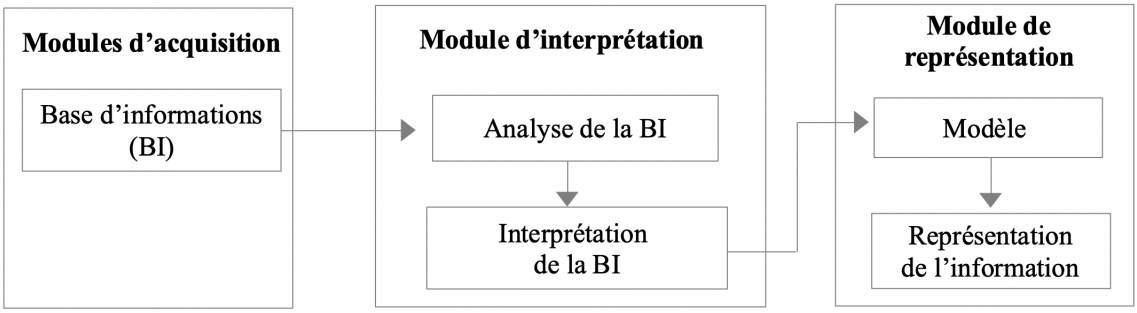

FIGURE 2.2 – Agrégation des données selon le contexte

Modèle de traitement des données

Finalement, le modèle de traitement des données est exploité par l'expert pour fournir une application sensible au contexte de l'expertise. Durant cette phase va s'effectuer l'implémentation d'un ensemble de mécanismes d'adaptation prévus pour donner suite

aux changements de contexte de traitement pour la prise de décision. Ces mécanismes peuvent être, par exemple, un ensemble de règles d'adaptation. Les changements contextuelles vont affecter l'analyse, la description et la représentation des connaissances.

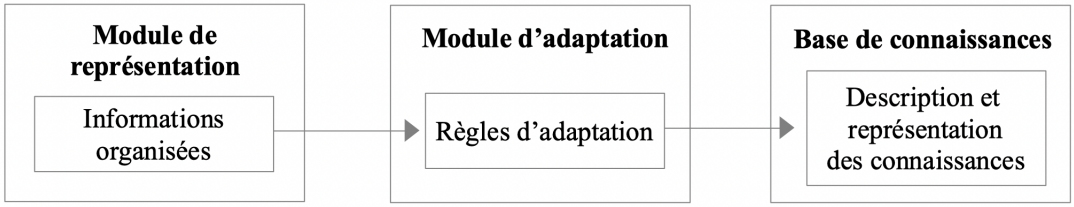

FIGURE 2.3 – Traitements des données

La démarche d'analyse de l'expertise humaine débute ainsi par l'acquisition des données par les outils physiques et logiques pour produire des connaissances et une représentation de ces dernières. La représentation des connaissances est le résultat de la transformation des observations du monde réel en une représentation numérique et logique pour la machine.

2.2 La représentation des connaissances

Selon Guarino (1995), une représentation des connaissances permet de dénoter des objets et de décrire les relations entre eux. Pour Levesque (1986), il s'agit d'écrire une représentation d'une partie du monde de telle façon qu'une machine puisse parvenir à de nouvelles conclusions sur l'environnement réel en manipulant cette représentation.
Les premiers travaux dans le domaine de gestion des connaissances datent de 1980 et les premiers résultats notables de 1985 (de Villaseñor, 1989 ; Chandrasekaran, 1987). Le but de ces travaux était d'obtenir une définition sur la manière de procéder pour recueillir les connaissances d'un système expert. Ces travaux se sont basés sur une approche cognitiviste, en restant au plus près de la démarche de raisonnement d'un expert. D'autres recherches, comme ceux de Aussenac-Gilles (2005) se sont inspirées des approches venant de l'intelligence artificielle et de plusieurs travaux convergeant vers l'acquisition de connaissances. En 1982, Newell (1982) a proposé une base théorique sur les recherches faites à partir de modèles conceptuels, en proposant l'analyse d'un système à base de

connaissances avec un « niveau ». Ce « niveau » de connaissances se trouve au-dessus du « niveau » formel dans une description en couches de plus en plus abstraites du fonctionnement des systèmes informatiques. Le système, perçu comme un agent rationnel par un observateur, y est spécifié en matière de buts pour organiser les heuristiques mises en œuvre par un système expert (de Villaseñor, 1989). Newell (1982) a aussi soutenu la notion de modèle conceptuel centré sur le niveau de connaissance « Knowledge level » et considéré comme une représentation à part entière. Enfin, Newell (1982) souligne les limites de la représentation de connaissances en insistant sur le fait que les connaissances ne peuvent être vues autrement que comme le résultat d'un processus d'interprétation s'appliquant à des expressions symboliques. Parmi les systèmes de représentation de connaissances on peut citer : Generic Tasks, Role Limitin Methods, TEIRESIAS, KADS.

« Generic Tasks » de Chandrasekaran (1987) sont des tâches génériques et primitives qui englobent des règles de production. Ils sont utilisés pour la description des buts et des raisonnements produits par un système à base de connaissances. Ses structures organisent les règles en fonction des buts au moment de la conception du système. Définies à un niveau générique, elles sont propres à un type de problématique pour guider un dialogue, elles sont propres à un type de problème et peuvent être utilisées dans différents domaines pour les mêmes tâches. Des travaux, en particulier « Components of Expertise »(Steels, 2002) et KADS, ont développé ces deux idées : rendre explicites des buts pour spécifier le système et représenter le raisonnement sous forme d'arbre de tâches [3] ; définir des blocs génériques réutilisables pour de nouveaux systèmes.

« Role Limitin Methods » est un ensemble de méthodes définies par l'équipe de J. McDermott au M.I.T. [4] Beaucoup de logiciels d'organisation de connaissances ont été définis selon ce principe (Marcus, 1988). Ils exploitent une représentation explicite de haut niveau de la résolution de problème pour guider un dialogue avec l'expert et l'enrichissement de la base de règles. Les règles jouent des rôles prédéfinis dans la résolution de problème, d'où le nom de l'approche. Ces systèmes ont permis l'établissement de plusieurs résultats qui ont été ensuite repris (la méthode de résolution de problème peut être explicitée, elle peut guider l'organisation des connaissances du domaine et les règles). Ces méthodes de

3. l'arbre des tâches correspond à la tâche d'exécution du protocole expérimental
4. Massachusetts Institute of Technology, est un institut de recherche dans les domaines des sciences et technologies, située aux états-unis.

résolution de problème sont adaptées à des types de problèmes et ne conviennent pas à tous. Enfin, ces méthodes sont des méthodes de l'intelligence artificielle, qui caractérisent l'algorithme de parcours des règles plus que la manière de raisonner de l'expert.

TEIRESIAS de R. Davis (1979) est un module de transfert d'expertise qui a été associé au premier système expert, MYCin, pour la correction des règles. Le système dialogue avec l'expert à partir des erreurs signalées pour guider le repérage des règles ayant conduit à cette erreur. TEIRESIAS valide l'approche d'une acquisition des connaissances interactive, s'appuyant sur le système d'inférence et s'adresse directement à l'expert du domaine.

Vogel (1988), à travers KOD, propose une méthode qui permet de définir des modalités d'entretien avec les experts et d'une analyse linguistique de leurs paroles, pour la construction d'un modèle cognitif qui sera exploitable dans un système expert. Vogel met l'accent sur la nécessité de comprendre des mécanismes cognitifs des experts et cherche à ce que le raisonnement produit par ce système corresponde à celui de l'expert. Le modèle cognitif est vu comme une représentation intermédiaire entre le langage naturel, un moyen d'expression des connaissances, et le système opérationnel. Enfin, Vogel a été un des premiers à proposer que l'opérationnalisation suive le paradigme objet et non uniquement à base de règles.

ETS de Boose (1986) devenue Aquinas (Boose et al., 1989) est un système qui s'appuie sur des travaux en psychologie pour proposer une technique de classification de concepts et de valeurs. Cette classification est appelée « les grilles de répertoires ». Elle propose une interface graphique d'expression des connaissances, puis de classification pour l'établissement des corrélations entre valeurs et concept, mais aussi des corrélations entre les concepts (Kassel, 2018). Ces corrélations débouchent sur la définition de nouvelles règles de production. Un des mérites de ce système est de susciter un questionnement sur des connaissances inattendues à acquérir à partir de concepts qui semblent à priori éloignés. Ce qu'on peut retenir de ce modèle est la notion de « Mediating Representation », c'est-à-dire l'intérêt de la construction d'un modèle qui s'appuie sur des représentations intermédiaires (sous une forme graphique simple, favorisant l'expression et l'interprétation des connaissances par l'expert).

KADS est une méthode d'analyse et de modélisation des connaissances. Il propose une modélisation de l'ensemble d'une base de connaissances via un cycle d'acquisition de

connaissances. Les recherches préalables à la première version de KADS ont rapidement été identifiées comme une proposition prometteuse, car intégrant les différentes contributions et réflexions précédentes. Les premiers travaux de Wielinga et al. (1992) posent le problème de l'acquisition des connaissances dans les mêmes termes.
Dans cette thèse, nous avons décidé d'élaborer une représentation des connaissances pour lever les verrous d'hétérogénéité des données qui ont été collectées par le logiciel Retaily après les enquêtes mystères. En effet, nous avons vu en introduction du manuscrit que la problématique majeure qui complexifie le traitement des données d'enquêtes mystères est liée à l'hétérogénéité de ces dernières, car les données collectées par le Logiciel Retaily durant les enquêtes proviennent de différentes sources (e.g. questionnaires, données des points de vente, etc.). Pour réaliser la représentation de ces connaissances, il est nécessaire en amont d'extraire ces connaissances à partir des données acquises.

2.2.1 Extraction des connaissances à partir des données

L'extraction de connaissances à partir de données est un processus qui se déroule suivant une suite d'opérations, des traitements informatiques permettant de transformer les données en informations puis en connaissance. Avant de parler de ces opérations, il est important de présenter les concepts : donnée, information, connaissance et savoir.

Donnée, information, connaissance et savoir

L'expertise humaine s'appuie sur la capacité à traiter des données, à analyser des informations et à représenter des connaissances. De ce fait, il est primordial de présenter la différence entre les concepts de donnée, information et connaissance. .
En prenant, référence des définitions de Zeleny (1987) ; Ackoff (1989) ; Burton-Jones (1999) ; Davenport et al. (1998) ; Rowley (2007), on peut présenter les concepts de donnée, information, connaissance et savoir comme suit :

— une donnée est un fait ou une observation, elle correspond à une « propriété d'un objet, d'un évènement ou de son environnement », assimilable à une entité non interprétée. En l'absence de contexte et d'interprétation, les données n'ont aucune

signification ;

— l'information est le résultat de l'interprétation d'un ensemble des données, notamment des données organisées, structurées. En prenant place dans un contexte spécifique, elle devient porteuse de sens. D'après (Ackoff, 1989), le sens, qu'il soit utile ou non, est établi par une connexion relationnelle ;

— la connaissance est liée à l'être humain contrairement aux données et aux informations. Elle est internalisée par un individu qui l'interprète sur la base de son expérience, de ses observations, il s'agit de la synthèse d'un ensemble de sources d'informations. Elle permet de répondre à la question « comment » ;

— Le savoir est un « processus extrapolatif, non déterministe et non probabiliste » ; (Ackoff, 1989). Il permet d'obtenir une facilité à la compréhension en faisant appel au jugement issu de types particuliers de programmations humaines tels que l'éthique et les codes moraux.

Ce spectre du savoir peut également être articulé le long d'un axe temporel. Alors que les trois premières catégories (donnée, information et connaissance) concernent le passé de ce qui a été le savoir comme une projection de l'avenir (voir Figure 2.4).

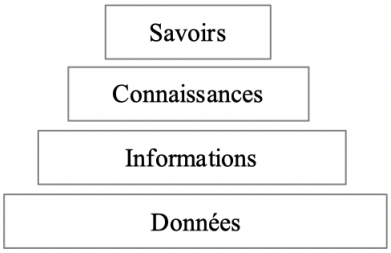

FIGURE 2.4 – Des données aux savoirs

Dans cette thèse, les questionnaires d'enquête mystère remplis contiennent des données articulées de sorte à fournir des informations. En mettant ces questionnaires en relation, on peut regrouper les informations et en extraire des connaissances. Si on a la possibilité de représenter et de classifier ces connaissances formellement, on peut espérer atteindre le savoir en favorisant la compréhension des connaissances.

Des données aux informations

Selon Stenmark (2002) l'information est l'ensemble de toutes les données externes aux personnes, communiquées oralement ou médiatisées par le biais de documents.

A l'opposé de la donnée, l'information possède une signification obtenue à partir de l'interprétation d'une ou plusieurs données. Ces données constituent la matière première à partir de laquelle sont produites les informations (Leleu-Merviel & Useille, 2008). Cette relation entre la donnée et l'information peut s'exprimer par la formule suivante :

$$Information = donnée(s) + signification \tag{2.1}$$

Par exemple, l'interprétation de l'ensemble des données d'achats effectués par un client sur un site de vente peut constituer une information. Néanmoins, la différence entre une donnée et une information peut varier en fonction de l'opinion. En effet, une donnée ne possédant pas de sens pour une personne, peut-être une information pour une autre personne.

Dans cette thèse, nous assimilons les données aux constats ou à des faits décrivant une évaluation d'un point de vente par un client mystère. Une fois que ces données sont mises en relation et interprétées par une personne, ces dernières deviennent des informations. L'interprétation des données prend en compte les expériences passées et les connaissances acquises par des prédictions. Cette étape d'interprétation est importante pour l'apprentissage, car les informations obtenues sont ensuite traitées pour construire des connaissances. Nous verrons qu'un SR interprète des informations pour faciliter la construction des connaissances, mais également pour faire des prédictions.

Des informations aux connaissances

L'analyse des informations par une personne consiste à transformer celles-ci en connaissances internes dans sa mémoire, ce qu'on peut aussi appeler « apprentissage ». Selon Stenmark (2002), la connaissance est le résultat de toute construction mentale internationalisée par un individu à partir d'informations qu'il obtient. La représentation des connaissances est le processus inverse par lequel une personne produite des informations utilisables par d'autres personnes, en utilisant un système de représentation, ce qu'on peut appeler « extraction des connaissances ».

Au début des années soixante, l'évolution de l'informatique a amené la création des systèmes de représentation des connaissances tels que les cartes conceptuelles, les réseaux sémantiques, les schémas ou cadres sémantiques, des modèles entités-relations, des modèles de flux d'information et des modèles orientés objet. Le but de ces systèmes est de proposer un langage formel, souvent graphique, permettant de faire une représentation des connaissances qui se « cachent » dans les informations, par exemple, deux textes, le premier en français et le deuxième en anglais, la traduction de l'un vers l'autre, seront représentés de la même manière par un modèle des connaissances regroupant les concepts dont traitent ces textes et les relations entre ces concepts. De même, un ordre décrivant un processus de travail et un texte écrit décrivant la même suite d'opérations seront représentés de la même manière dans un système de représentation des connaissances. On parle alors de représentation du sens d'un document ou de représentation sémantique. Une représentation sémantique est une image du modèle mental des connaissances dégagées des particularités du format choisi pour présenter les informations.

Connaissances au savoir

Le savoir est un « *ensemble de connaissances d'une personne ou d'une collectivité acquise par l'étude, par l'observation, par l'apprentissage et/ou par l'expérience* »[5]. Selon Bouchereau (2020) cette définition présente le savoir en deux conceptions : le savoir pour une personne, ce que nous avons défini comme étant une connaissance, et le savoir collectif, élaboré par les chercheurs et scientifiques. Otlet (1934) formalise dans ses travaux le processus de construction des savoirs, leur encadrement par la science et leur ordonnancement selon des classifications. L'Homme observe le Monde et traduit sa pensée à travers des ouvrages qui sont classés et organisés par la suite dans des encyclopédies. Ce processus est une illustration de construction du savoir sur la base des observations du Monde. Néanmoins, c'est une construction humaine et exécutée dans un cadre socioculturel particulier. En effet, selon Morin (2008), les savoirs sont le résultat d'une activité collective à laquelle prennent part des personnes qui observent et décrivent le réel à travers les cadres de pensée de leur époque. Ainsi, tout savoir « *subit une détermination sociologique. Il y a dans toute science, même la plus physique, une dimension anthropo-sociale* » (Morin,

5. Source : https ://www.cnrtl.fr/definition/savoir. Consulté le 7 juillet 2021.

2008).

Traitement des informations et des connaissances

Pour Stenmark (2002), le traitement de l'information consiste à faire circuler au sein d'une organisation, les informations du domaine dans lequel celle-ci opère, ainsi que des informations internes et externes sur les biens et services produits par cette dernière et sur ses processus de travail. Selon Akdag et Khoukhi (1994), certaines de ces informations sont retenues et traitées pour constituer la mémoire de l'organisation sous la forme de base de données ou de base de documents, lesquelles sont utilisées pour produire bien et services.

Le traitement des connaissances ajoute l'analyse à ce processus classique, c'est-à-dire la transformation des informations en connaissances chez le personnel par l'apprentissage (Akdag & Khoukhi, 1994). Selon Akdag et Khoukhi (1994), le processus inverse est l'extraction et la représentation des connaissances pour les intégrer dans la mémoire de l'organisation sous la forme accessible à l'ensemble du personnel.

En effet, le composant principal d'un système de traitement des connaissances est la base de connaissances qui regroupe une représentation standard des connaissances du domaine. Cette représentation des connaissances sert à faire une structuration du contenu des documents ainsi que des données qui se trouvent dans la mémoire de l'organisation. Ainsi, la base de connaissances sert à référencer de manière intégrée les documents et les informations qu'elle contient (Akdag & Khoukhi, 1994). Elle permet aussi d'effectuer une recherche sur les informations et de repérer des documents alternatifs pour inclure des connaissances qui n'ont pas été prises en compte.

Selon Akdag et Khoukhi (1994), la représentation de connaissances se limite au traitement des documents. Cette approche conduit à des activités non structurées et non reliées de création, d'importation, de capture, de recherche et d'utilisation des connaissances. Des composants de connaissances peuvent exister dans un large type de format de documents ou dans des connaissances implicites détenues par des personnes, mais encore exprimées dans la mémoire de l'organisation. Au niveau de la recherche d'informations, ces dernières sont référencées de manière syntaxique par un moteur de classement, remplaçant le référencement direct par un modèle de connaissances. Ainsi, la

problématique est le suivi des associations entre concepts pour repérer des documents reliés ou complémentaires. Cette association donne la possibilité à des agents informatiques d'effectuer une recherche et de trouver des documents par leur sémantique, par leur contenu, plutôt que par leur syntaxe, ou par leur forme. Les documents accessibles par internet sont associés entre eux. La représentation des connaissances de cette association est à la base du web sémantique.

Le processus de traitement des connaissances comporte quatre grandes étapes :

1. la création ou l'importation des informations sous la forme de textes, images fixes et animées, modèles graphiques, etc ;
2. le référencement sémantique des informations et des documents, indexées et décrites par leur attribut, par des métadonnées, par leur position dans des classifications, en utilisant des ontologies et généralement dans le cadre d'un système de représentation des connaissances ;
3. la recherche et l'accès aux informations en identifiant d'abord des éléments de connaissances dans la base de connaissances, plutôt que des données dans une base documentaire. Cette dernière étant référencée en fonction des connaissances, on la retrouvera en utilisant les métadonnées comme base de la recherche. On pourra aussi utiliser le modèle des connaissances ou l'ontologie de référencement sémantique pour la navigation dans les informations, effectuer des recherches, obtenir l'accès à des pages web, des documents, des personnes ;
4. l'utilisation des informations dans le cadre de processus de travail par la segmentation, l'agrégation, l'annotation et l'intégration des informations dans les nouvelles ressources informationnelles ou documents qui seront à leur tour référencés de manière sémantique et intégrés dans le processus de traitement des connaissances.

Une base de connaissances est composée d'une partie terminologique, qualifiée de « TBox » (Terminological Box), et d'une partie assertionnelle, qualifiée de « ABox » (Assertional Box). La « ABox » décrit les individus et leurs relations (quel individu appartient à quel concept nommé, quel individu est lié à quel autre à travers quel rôle). La mise en place d'une base de connaissances nécessite la description d'une ontologie qui correspond au domaine du marketing, des règles qui viennent compléter cette dernière (Rbox) et des annotations qui

sont des données décrivant des données (appelées méta- données) (Gandon et al., 2012). Traditionnellement, en ingénierie des systèmes, la représentation des données s'effectue sur la base de méthodes traditionnelles comme : Merise[6], OMT[7], UML[8], etc. Ces représentations traditionnelles sont issues de modèles d'association, d'objet, de réseaux, graphes ou encore modèles de flux. Ces modèles permettent de favoriser un continuum de la définition des besoins des clients jusqu'au système développé et exploité. C'est une approche qui conduit à l'analyse, à la conception et à l'implémentation des systèmes informatique. Une nouvelle approche basée sur la réutilisation de composants commence à émerger.

Avec le passage des données aux connaissances, le mode de représentation est conservée, mais avec des modèles permettant de gérer des connaissances. Ces modèles sont construits autour des technologies sémantiques. Ainsi la représentation des connaissances est incontournable pour les systèmes à base de connaissances.

Fouille de données

Selon Fayyad et al. (1996), la fouille de données[9] consiste à donner un sens aux grandes quantités de données, d'un certain domaine, capturées et stockées massivement par les entreprises. En effet, la vraie valeur n'est pas dans le fait d'acquérir et de stocker des données, mais plutôt dans notre capacité à en extraire des connaissances utiles et à trouver des tendances et des corrélations intéressantes pour appuyer les décisions des décideurs d'entreprises et des scientifiques. Cette extraction fait appel à un ensemble de méthodes, d'algorithmes et d'outils provenant de statistiques, d'intelligence artificielle, de bases de données, etc.

Le développement récent de la fouille de données est lié à plusieurs facteurs (Xuan et al., 2021) :

6. Merise est une méthode d'analyse et de conception dans le cadre de gestion de projet informatique. Il a été très utilisée dans les années 1970 et 1980 pour l'informatisation massive des organisations.

7. ≪ Object modeling ≫ est une technique de modélisation et de conception pour la programmation orientée objet. Elle a été conçue en 1991 par James Rumbaugh.

8. ≪Unifed modeling language est un langage de modélisation graphique à base de pictogramme conçu pour fournir une méthode normalisée de visualisation de la conception d'un système.

9. Plus connu sous le nom d'exploitation de donnée, prospection de données ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données sur la base de méthodes automatique ou semi- automatique.

— une puissance de calcul importante est disponible ;
— le volume des bases de données augmente énormément ;
— l'accès aux réseaux de taille mondiale, ayant un débit sans cesse croissant, qui rendent le calcul distribué et la distribution d'information sur un réseau d'échelle mondiale viable ;
— la prise de conscience de l'intérêt commercial pour l'optimisation des processus de fabrication, vente, gestion, logistique.

Il importe de ne pas faire comme si toutes les données ont une valeur connue, et encore moins une valeur valide ; il faut donc gérer des données dont certains attributs ont une valeur inconnue ou invalide ; on dit que les données sont bruitées. La simple élimination des données ayant un attribut dont la valeur est inconnue ou invalide pourrait vider complètement la base de données. On touche le problème de la collecte de données fiables qui est un problème pratique très difficile à résoudre. En fouille de données, il faut faire avec les données dont on dispose sans faire comme si l'on disposait des valeurs de tous les attributs de tous les individus.

Cependant avant de tenter d'extraire des connaissances utiles à partir de données, il est important d'avoir une procédure bien claire et d'en comprendre la démarche dans sa globalité. Il est nécessaire de connaître les algorithmes d'analyse de données et savoir les appliquer sur des données en mains pour la bonne conduite d'un projet de fouille de données. Une application aveugle des méthodes de fouille de données sur les données en main peut mener à la découverte de connaissances incompréhensibles, voire inutiles pour l'utilisateur final (Fayyad et al., 1996). Pour cette raison l'extraction de connaissances et la fouille de données ont été rapidement organisées sous forme d'un processus appelé processus d'extraction de connaissances à partir de données (ECD). Ce processus se présente comme un processus complexe, non trivial, composé de plusieurs étapes itératives, et nécessitant une interactivité permanente de la part de l'utilisateur expert. Le processus constitue une feuille de route à suivre par les praticiens lors de la planification et la réalisation des projets d'extraction de connaissances à partir de données. Pour Fayyad et al. (1996), l'extraction de connaissances de données émerge comme un domaine à part entière, sans remettre en cause ses origines, pour intégrer de nouvelles problématiques. On peut même annoncer, sans craindre de critiques, en ingénierie d'extraction et de ges-

tion de connaissances disposent actuellement de ses propres modèles, méthodologies et langages (Charlet, 2002).

Le processus de l'ECD comporte des étapes de prétraitement qui ont lieu avant la fouille de données proprement dites. Le prétraitement porte sur l'accès aux données en vue de construire des corpus de données spécifiques. Le prétraitement concerne la mise en forme des données entrées selon leur type (numérique, symbolique, image, texte, son), ainsi que le nettoyage des données, le traitement des données manquantes, la sélection d'instances. Cette première phase est cruciale, car la mise au point des modèles de prédiction va dépendre du choix des descripteurs et de la connaissance précise de la population. L'information nécessaire à la construction d'un bon modèle de précision peut être disponible dans les données, mais un choix inapproprié de variables ou d'échantillons d'apprentissage peut faire échouer l'opération.

Pour Fayyad et al. (1996), le traitement de l'ECD est appelé « niveau analyse », car les données provenant des bases de données alimentent les entrepôts de données qui seront exploitées en ECD. Le traitement d'ECD se déroule généralement en quatre phases, sous la supervision d'un expert :

1. acquisition de données a pour but de cibler l'espace de données qui va être exploré ;
2. permet le prétraitement et la mise en forme des données. Toutes Les données issues de l'entrepôt ne sont pas nécessairement exploitables par des techniques de fouille de données ;
3. fouille de données, une confusion entre les termes « fouille de données » et « l'extraction de connaissances à partir de données » existe depuis un certain temps. En effet pour beaucoup de chercheurs et praticiens le terme « fouille de données » est utilisé comme un synonyme de l'extraction de connaissances de données en plus d'être utilisé pour décrire l'une des étapes du processus de l'extraction de connaissances de données ;
4. analyse et mise en forme des connaissances.

Ils soulignent que l'extraction de connaissances s'effectue sur des tables bidimensionnelles, appelées « data-marts », et fait appel à trois grandes familles utilisant des méthodes statistiques, d'analyse des données, de la reconnaissance de formes ou de l'ap-

prentissage automatique. Ces méthodes sont utilisées ou présentées comme faisant partie des outils de la fouille de données :

— les méthodes de structuration uni-, bi- et multidimensionnelles. Numériques, pour la plupart, ces méthodes sont issues de la statistique descriptive et de l'analyse des données, ainsi que des techniques de visualisation graphique dont certaines font appel à la réalité virtuelle et à des métaphores calquées sur le modèle mental humain ;

— les méthodes de structuration qui regroupent toutes les techniques d'apprentissage non supervisées et de classification automatique provenant des domaines de la reconnaissance de formes, de la statistique, de l'apprentissage automatique et du « connexionnisme » [10] ;

— les méthodes explicatives dont le but est de relier un phénomène à expliquer à un phénomène explicatif. Généralement mises en œuvre en vue d'extraire des modèles de classement ou de prédiction, ces méthodes descendent de la statistique, de la reconnaissance de formes, de l'apprentissage automatique et du « connexionnisme », voire du domaine des bases de données dans le cas de la recherche de règles d'association.

En dehors du domaine des statistiques (Lecomte & Quatrini, 2010 ; Fayyad et al., 1996) citent dans leurs travaux des algorithmes de recherche de règles d'association dans les grandes bases de données. Les premiers algorithmes proposés dans le domaine des statistiques ont satisfait des statisticiens en raison de la naïveté du matériel méthodologique qui était alors utilisé.

2.2.2 Web Sémantique et ontologie

Plusieurs raisons peuvent causer la difficulté de compréhension d'un dialogue, empêchant de faire comprendre à l'autre ce que l'on désire exprimer. Bien souvent, nous passons par de petits dessins pour mieux exprimer les choses, ces petites représentations plus ou moins formelles permettent un accord sur l'interprétation à adopter pour nous comprendre (Fayyad et al., 1996).

10. Approche de modélisation basée sur l'utilisation des réseaux neuromimétiques

Dans la phrase « Loïc fait la cuisine », plusieurs interprétations peuvent être faites : Loïc prépare à manger, Loïc monte les meubles de la cuisine ou encore Loïc peint les murs de la cuisine (la pièce de la maison où l'on prépare à manger). Toutes ces interprétations sont dues aux différents sens que l'on peut inférer à partir des termes utilisés, en fonction du contexte. Si dans notre exemple, nous définissons le terme cuisine comme la pièce de la maison dans laquelle on prépare à manger, nous restreignons alors les interprétations possibles.

Pour éluder le plus possible les ambiguïtés de compréhension, nous utilisons naturellement une convention entre les participants au dialogue, passant par un formalisme qu'il soit graphique ou linguistique. C'est dans ce formalisme que les ontologies prennent part. Par conséquent, dans ce qui suit, nous allons détailler la relation entre le web sémantique et les ontologies. Tout d'abord, nous introduisons le web sémantique. Ensuite, nous allons présenter une définition, la modélisation et la formalisation des ontologies. La structure d'une ontologie, la taxonomie des domaines et la classification des ontologies.

Web Sémantique

Selon Berners-Lee et al. (2001), le Web tel qu'on le connaît aujourd'hui est principalement syntaxique c'est-à-dire que la structure des ressources y est bien définie, mais que le sens de leur contenu reste inaccessible aux ordinateurs. L'interprétation des contenus nécessite donc une intervention humaine. Le Web 3.0 ou le Web sémantique a pour ambition de lever cette contrainte en facilitant l'accès aux ressources du Web aussi bien par l'homme que par la machine, grâce à la représentation sémantique de leurs contenus.

Les technologies mobilisées dans ce cadre pour l'encodage des données ont fait l'objet depuis 1998 d'un ensemble de recommandations du W3C (World Wide Web Consortium) [11] qui ont été peu à peu amplifiées et actualisées. Le W3C propose ainsi le schéma en couches du web sémantique, familièrement appelé « layer cake », ici dans une version de comment ces technologies s'articulent et signalent la volonté d'aboutir à une architecture cohérente(voir figure 2.5).

11. Le World Web Consortium est un organisme de standardisation créé en octobre 1994. Il est charge de promouvoir la compatibilité des technologies du web telles que RDF, SPARQL, CSS, HTML5, XSL, etc. http ://www.w3c.org

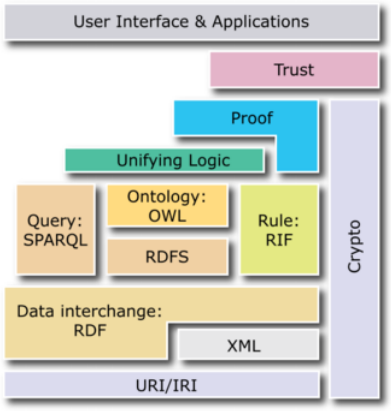

FIGURE 2.5 – Le « layer cake » du web sémantique

Selon Berners-Lee et al. (2001), le web sémantique est une extension du web actuel dans le lequel on donne à une information un sens bien défini pour permettre aux ordinateurs et aux individus de travailler en coopération . Ainsi le W3C met en place les recommandations suivantes :

— la description du web par des classifications précises, par le biais d'ontologies exploitables par les machines et compréhensibles par les humains ;

— l'utilisation d'un langage commun pour exprimer les ontologies et décrire des annotations utilisant leurs termes ;

— la création de moteurs de raisonnement permettant d'inférer sur les annotations d'après les axiomes déclarés dans les ontologies.

La technologie du web sémantique est une infrastructure permettant une formalisation des connaissances qui va bien au-delà du contenu informel du web classique.

L'objectif principal du web sémantique est de procurer aux utilisateurs finaux des services plus intelligents basés sur l'utilisation par la machine de connaissances représentées en exploitant des ontologies et des bases de connaissances (Berners-Lee et al., 2001). L'architecture qui a été proposée par Tim Berners-Lee, repose sur une pyramide de langages pour représenter des connaissances sur le web (voir Figure 2.5).

Ontologie

L'ontologie se définit sur la base du concept de réseau sémantique (Rastier, 1995 ; Desclés, 1987 ; Chantrain, 2017). Quillian (1967) est cité comme une référence incontour-

nable, pour avoir utilisé le concept de réseau sémantique et avoir modélisé le fonctionnement de la mémoire (Geller, 2009). Selon Geller (2009), le but de ce concept est de décrire la réalité sous forme de réseaux ou encore de graphes. Pour Geller (2009), ce réseau est constitué de nœuds qui représentent les différents concepts qui sont reliés par des arcs. Ces arcs expriment la relation entre les concepts. Les arcs et les nœuds sont le plus souvent étiquetés. Dans un premier temps les objets sont associés, ensuite les relations entre eux sont définies pour obtenir la structure (objet, relation). Les relations les plus spécifiques de ce type de réseau sont les relations (sorte-de). D'autres relations qui sont utilisées par exemple (instrument pour) ou encore (a pour partie), etc. La relation (sorte-de) permet de définir l'une des notions les plus essentielles des réseaux sémantiques et de tout autre formalisme qui intègre une logique correspondant à la déduction des propriétés par héritage. Cet héritage est basé sur la transitivité de la relation (sorte-de). L'idée de l'héritage se positionne comme le moyen le plus économique sur le fait de rattacher une caractéristique commune à un lot de concepts au niveau le plus élevé de la hiérarchie. Dans ce contexte, il n'est pas nécessaire de stocker l'ensemble des propriétés en mémoire pour chaque concept, car cela augmente le pouvoir d'expression de la recherche d'informations, mais également de la modélisation. L'utilisation du concept de réseau sémantique a migré vers la représentation des connaissances en particulier dans le cadre de la modélisation des ontologies. Une ontologie est un réseau sémantique qui renferme un ensemble de concepts décrivant un domaine ou une partie de ce dernier. Elle est utilisée dans le but de définir ou encore de représenter un fait ou un raisonnement. Son utilisation a connu une croissance significative avec les outils du web sémantique. Cet outil qui sert de référence pour la communication entre les ordinateurs et entre les humains et les ordinateurs sur la base de la définition du sens des objets. L'ontologie permet la réutilisation et le partage des données. Selon Gruber (1993), les techniques d'interopérabilité basées sur les communications peuvent être définies sur trois niveaux : le protocole de communication des agents, la spécification du vocabulaire partagé et le format de représentation du langage (Gruber, 1995 ; El Bouhissi et al., 2020).

Pour Gruber (1993), la catégorisation de la communication peut se faire en trois parties : communication système-système, humain-système et humain-humain. Ces parties ont chacune des caractéristiques avec des problèmes qui peuvent être résolus par une onto-

logie. Ainsi, on peut dire que l'ontologie permet d'uniformiser le langage d'échange entre les différents agents, de faire une comparaison des différents systèmes, de rendre le vocabulaire standard, de synthétiser les connaissances du domaine, de spécifier les contextes et de structurer la connaissance pour simplifier l'analyse. Selon Ikeda et Stephens (1998), on peut utiliser l'ontologie pour supporter la recherche d'informations en mathématique par exemple pour apporter une base théorique à la physique. Pour Razmerita et al. (2003), se baser sur les ontologies et sur le web sémantique permettrait d'avoir une ouverture sur de nouvelles possibilités et des défis à la conception d'un ensemble de systèmes adaptatifs en rendant possible la modélisation du profil utilisateur.

En effet, les ontologies sont devenues incontournables dans le contexte d'adaptation de l'information à l'utilisateur. Elles donnent la possibilité de construire des modèles de connaissances qui sont utilisables pour la modélisation à la fois du domaine et des utilisateurs. L'adaptation du web à l'utilisateur pourrait favoriser le développement du web et le faire migrer vers un web adaptatif.

Le modèle de domaine nous permet de faire une représentation du contenu en fonction de la connaissance générale du domaine. Ce modèle est composé d'une ontologie du domaine qui est une sous partie de l'ontologie générale du système, l'ontologie du domaine est orientée objet. Elle est utilisée pour la représentation du domaine sous une forme de base de connaissances.

Dans notre thèse, notre ontologie va faire une représentation du domaine du marketing. L'ontologie présente les concepts, les propriétés et les instances du domaine.

Les langages ontologiques du web sémantique

Le web est conçu comme un espace d'information, dont le but est d'être utile non seulement pour la communication humaine, mais aussi pour que les machines puissent participer et aider (Gandon et al., 2012). L'un des principales limites à cette situation est le fait que la plupart des informations sur le web sont conçues pour la consommation humaine et même si elles proviennent d'une base de données avec des descriptions bien établies, la structuration de ces dernières n'est pas évidente pour un artéfact computationnel externe exploitant ces données.

Le web sémantique est un « Web de données », ainsi le W3C soutient la création de tech-

nologies permettant de prendre en charge un « Web de données ». L'objectif du « Web de données » est de permettre de développer des systèmes capables de prendre en charge des interactions sur le web et de donner un rôle important à l'ordinateur dans son travail. Le terme « Web sémantique » est une vision du W3C du web des données liées. Les technologies du web sémantique permettent aux utilisateurs de créer leurs bases de données sur le web, de créer des vocabulaires et d'établir des règles de traitement des données. Les données liées reposent sur des technologies telles RDF [12], SPARQL [13], OWL [14] et SKOS [15]. Les recommandations du web sémantique tels que RDF, RDFs et OWL reposent essentiellement sur le graphe RDF. En effet, RDF est un langage du web sémantique qui permet de faire la description des données. Il décrit les propriétés des ressources ou les relations entre les ressources. Ce modèle est composé de ressources web comme les pages web, les images, etc. Chacune de ces ressources est identifiée par une URI et permet de faire des assertions, déclarer les ressources sous forme de triplet (sujet/prédicat/objet) (Gandon et al., 2012) (voir Figure 2.6).

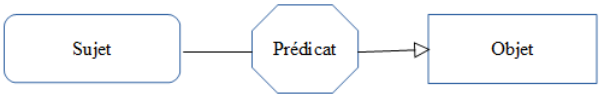

FIGURE 2.6 – Triplet

Le triplet sujet/prédicat/objet peut être assimile à un rapprochement avec le langage naturel et le triplet sujet/verbe/complément web. Comme le montre la figure 2.4, la composition de toute expression en RDF est une collection de triplets sous la forme (sujet, prédicat, objet). Chaque triplet est défini par un arc prédicat orienté du nœud source sujet vers le nœud destination objet. La globalité des triplets RDF représente un graphe orienté qui porte le nom de graphe RDF. RDF utilise des données de types ressources, des propriétés et des valeurs littérales pour la construction des triplets. En

12. Ressource Description Framework
13. C'est un langage de requête et un protocole qui permet d'ajouter, de modifier, de supprimer ou encore de rechercher des données RDF.
14. Web Ontology Langage est un langage de représentation des connaissances établit sur la base du modèle de données RDF. Il offre les possibilités de définir des ontologies web structurées.
15. Simple Knowledge Proganization System est une ensemble de langages formels qui permet de faire une représentation des thésaurus.

RDF, le sujet est obligatoirement un objet de type ressource, c'est à dire pouvant être référencé par une URI. Le prédicat quant à lui doit être de type propriété, identifié par une URI. Chacune des propriétés a un sens bien donné qui indiquera la sémantique de description. Les objets sont identifiables sous forme d'URI de ressource ou de chaîne de caractères (littéral).Par exemple :

— (« Loic Favory », est, une personne) ;
— (« Loic Favory », est président, Effet B) ;
— (Valentin, possède, PageValentin) ;
— (Valentin, son nom de site web est, http ://valentin.fr) ;
— (EffetB, son nom de site web est, http ://effetb.fr).

De notre exemple, on peut identifier les URI locales, les URI externes et les valeurs littérales. URI locales : Loic, EffetB, est président, son nom est URI externes : http ://valentin.fr, http ://effetb.fr Valeurs littérales : « Loic Favory » Dans l'exemple illustré sur la figure 7, ci-dessous, la représentation du graphe peut être traduit sous la forme de triplets :

— (Point de vente, est une, Entreprise) ;
— (Point de vente, a été créé, 7 septembre 2007) ;
— (Point de vente, est spécialisé, Automobile) ;
— (Automobile, est , Voiture sans permis).

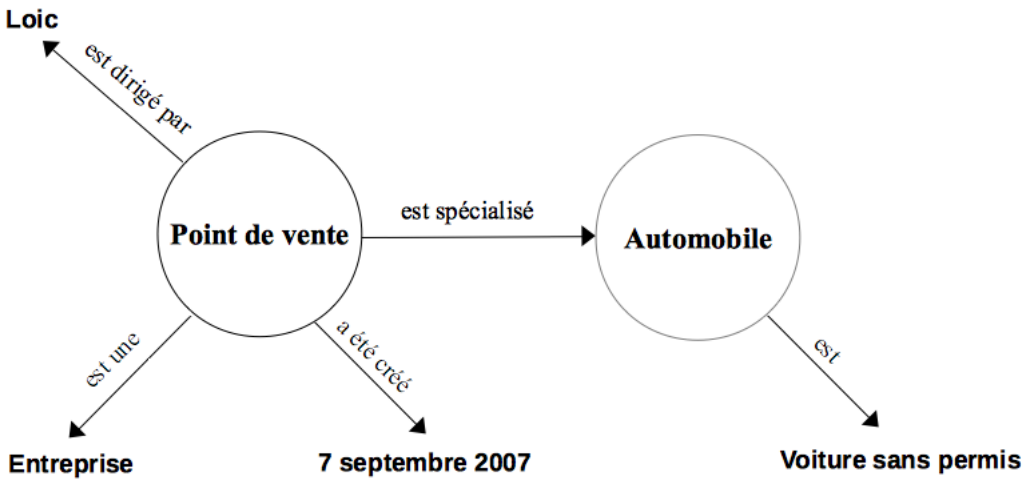

FIGURE 2.7 – Exemple de schéma RDF

Dans l'exemple on considère que Loïc, Point de vente, 7 septembre 2007, Voiture sans permis et Automobile sont des ressources. En RDF les relations est une et est dirigé par la catégorie Entreprise sont considérées aussi comme des ressources, car les concepts de ces derniers peuvent être décrits. Utiliser le terme Point de vente pour la description d'une entreprise créée le 7 septembre 2007 et qui s'est spécialisé à l'automobile est peut être une source de confusion, car le Point de vente peut être identifié ici comme deux entreprises différentes possédant la même appellation, l'une décrit comme l'entreprise qui est créée le 17 septembre 2007 et l'autre qui s'est spécialisée dans l'automobile. C'est pour cette raison que l'on fait appel à RDF qui possède un mécanisme d'identification des ressources. Cette identification assure qu'un identificateur est utilisé pour faire référence à une ressource et ce même identificateur peut être utilisé ailleurs ou dans un autre contexte pour référence unique. Les ressources en RDF peuvent être sérialisées en utilisant plusieurs syntaxes suivantes :

— la syntaxe RDF/XML, qui est une expression en XML [16] de données en RDF. C'est une syntaxe normalisée par le W3C ;

16. Extensible Markeup Language est un langage de description qui a pour but de formaliser des données textuelles. Il permet de faire la mise en forme de documents en utilisant des balises.

— les syntaxes spécifiques de type N3 [17], N-Triples [18] et Turtle [19] ;

— la syntaxe RDFa qui permet de faire une encapsulation des données RDF dans une page web.

Le modèles RDF est basé sur deux niveaux :

1. un niveau physique qui est composé de triplets et déclarations qui sont de types de base, des ressources, propriétés, déclarations et des types complexes comme les collections et les listes ;
2. un niveau schéma (RDFs) qui est composé de classes et de types de propriétés. RDFs est une extension de RDF. Il propose un modèle de description de vocabulaire RDF basé sur des classes et des propriétés. C'est un langage qui permet de décrire des langages RDF. Ce schéma permet de définir les types de ressources (livre, personne, etc) ainsi que leurs propriétés (diplôme, titre, auteur, etc.).

RDFS propose de l'information sur l'interprétation des déclarations RDF. Il permet aussi d'étendre RDF à la description d'ontologies par une hiérarchisation des classes et des propriétés (subClassOf, subPropertyOf), par la description du range and domain sur les propriétés. RDFS permet aussi de décrire des annotations (seeAlso, isDefinedBy, label, range, domain, member). Pour ses classes et propriétés ont peut citer :

— la classe Class qui est un ensemble regroupant plusieurs objets.

— la propriété subClassOf qui permet de définir une classe, sous-ensemble d'une autre classe ;

— la classe Resource qui représenta la classe parente ;

— la propriété range qui donne des indications sur le champs d'application d'une propriété ;

— la propriété domain qui favorise la spécification des classes auxquelles on peut affecter une propriété.

17. Notation 3 est langage d'assertion et logique qui est un sur-ensemble de RDF. N3 étend le modèle de données RDF en ajoutant des formules (les littéraux qui sont des graphes eux-mêmes), des variables, une implication logique et des prédicats fonctionnels, tout en fournissant une syntaxe textuelle alternative à RDF/XML.

18. Un format de stockage et de transmission de données. C'est un format de sérialisation en texte brut basé sur des lignes destinées aux RDF.

19. Une syntaxe de langage permettant une sérialisation non-XML des modèles RDF. Il offre des niveaux de compatibilité avec les formats N-Triples et Notation 3 existants, ainsi qu'avec la syntaxe à trois motifs de la recommandation proposée par SPARQL W3C.

Pour inclure la sémantique dans le processus de traitement de la recommandation, le calcul de la similarité sémantique est incontournable (Slimani, 2013).

2.2.3 Similarité sémantique

Définitions et notions

La similarité sémantique est au centre des recherches en science de l'information et traite les notions de taxonomie, de concept, d'arc, de distance et de profondeur d'un concept (A. N. Ngom, 2015 ; M. A. N. Ngom, 2018 ; Dudognon et al., 2010).
La taxonomie vise à établir une classification systématique des êtres vivants (Zargayouna, 2005), elle est utilisée pour représenter un ensemble des concepts unis par des relations hiérarchiques (Coenen-Huther, 2007). Cette forme de représentation est utilisée dans les systèmes à base de connaissances, notamment, les ontologies. Les ontologies sont illustrées sous forme d'arbre ou de taxonomie et la figure 3.1 qui suit est un exemple de représentation taxonomique d'une ontologie (Gómez-Pérez, 2004).

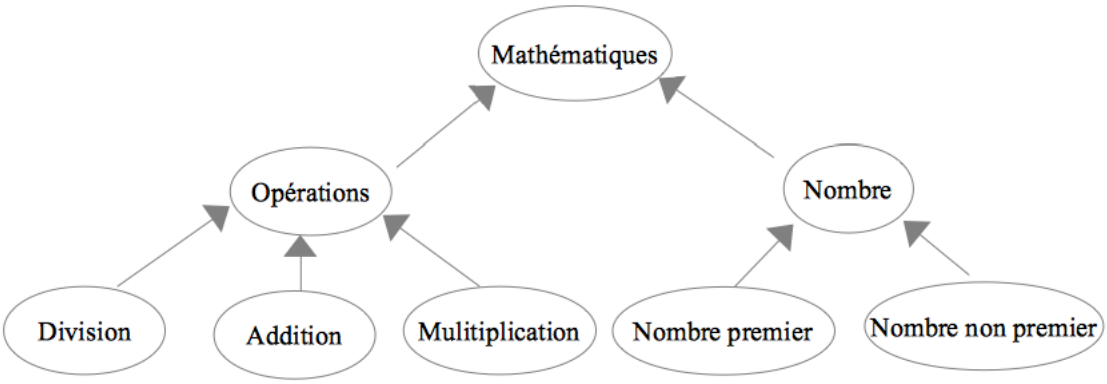

FIGURE 2.8 – Exemple de taxonomie

Un concept est une idée d'un groupe de personnes ou d'objets qui partage les mêmes caractéristiques (Dramé, 2014). Ses caractéristiques sont représentées dans l'ontologie par un ensemble de propriétés qui définissent une relation entre le concept et un domaine de valeurs bien précis (Niang, 2013). Un concept est généralement défini par un terme et a généralement une intension [20] et/ou une extension (Ogden & Richards, 1923). L'intension

20. Ensemble des caractères qui constituent un concept (par opposition à extension).

désigne la sémantique, c'est-à-dire, l'ensemble des attributs et propriétés qui définissent un concept. L'extension, quant à elle, désigne l'ensemble des objets qu'englobe un concept. Le Figure 2.9 est une illustration de ces éléments représentés dans un triangle sémantique (Ogden & Richards, 1923).

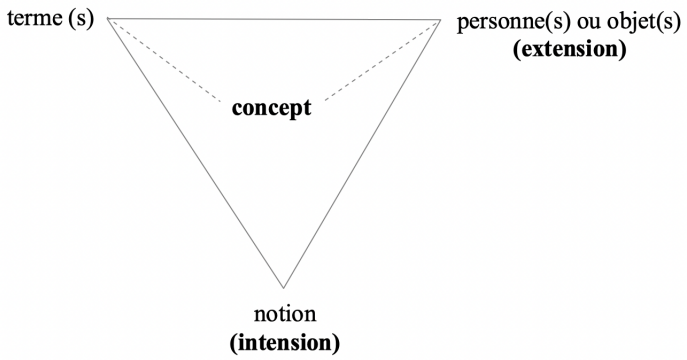

FIGURE 2.9 – Triangle sémantique

Par exemple, sur la Figure 2.8, « Nombre » et « Nombre premier » sont des exemples de concepts. Si nous prenons l'exemple du concept « Nombre premier », nous avons :

— son intension : un entier naturel qui admet deux diviseurs distincts entiers et positifs qui sont exactement 1 et lui-même ;

— son extension : tous les nombres qui suivent à cette définition.

Un arc permet de représenter un lien existant entre deux concepts dans une structure hiérarchique. Dans certaines ontologies particulières comme WordNet (Sussna, 1993), certains liens ont été définis, mais le plus utilisé est le lien (est-un) qui désigne une relation spécifique entre deux concepts. Lorsqu'un concept est spécifique à un autre, on dit qu'il est subsumé par celui-ci. Par exemple, dans la Figure 3.1, nous avons l'entité nombre qui est spécifique à l'entité math, alors nous pouvons dire que nombre est subsumée par math.

La distance représente quant à elle le plus petit nombre d'arcs qui sépare deux concepts dans une taxonomie.

La profondeur d'un concept ci dans une taxonomie est le niveau de ce concept par rapport à la racine de la taxonomie. La profondeur du concept c_i est notéeP_i. La profondeur totale d'une structure hiérarchique est la valeur maximale des profondeurs de l'ensemble de ces éléments.

Mesures de similarité sémantique

Les mesures de similarité sémantique sont des fonctions très utilisées dans les domaines des sciences de l'information parmi lesquels nous avons l'informatique, la Bioinformatique, le traitement automatique des langues naturelles, l'ingénierie des connaissances, etc. Elles permettent de déterminer la similarité entre des concepts qui n'ont aucune ressemblance syntaxique. Leurs utilisations se basent généralement sur une bonne organisation des concepts en structure hiérarchique grâce à l'utilisation des outils de représentation de connaissances comme les ontologies.

Depuis les années 90, plusieurs types de mesures ont été définis. Ainsi, nous pouvons classifier ces mesures en trois groupes Jiang et Conrath (1997) ; Elavarasi et al. (2014) ; R. Gupta et Singh (2017) :

- les mesures calculant la distance entre les concepts sur la base du nombre d'arcs qui les séparent ;
- les mesures basées sur la quantité d'information partagée par les concepts grâce à l'utilisation de la théorie de l'information ;
- les mesures hybrides utilisant la combinaison des deux groupes cités plus haut ou sur l'usage de diverses techniques.

Il existe de nombreuses mesures de similarité sémantique, avec des propriétés et des résultats différents. Dans cette thèse, nous souhaitons classer tous les concepts du domaine d'application par rapport à un concept central, c'est-à-dire un concept de référence ou encore la racine. Il s'agit donc pour nous de choisir la meilleure mesure de similarité sémantique, étant données ces contraintes et eu égard aux résultats des différentes mesures. Il paraît donc évident d'utiliser les chemins (suite d'arcs du graphe) pour mesurer la distance entre les concepts. Selon Rada et al. (1989), la démarche la plus intuitive. Il présente ainsi une mesure, dist (c_1, c_2), indiquant le nombre d'arcs minimum à parcourir pour aller d'un concept c_1 à un concept c_2 :

$$Sim(c_1, c_2) = \frac{1}{(1 + dist(c_1, c_2))} \tag{2.2}$$

D'autres mesures utilisent la notion de plus petite généralisant commun, c'est-à-dire le généralisant commun à c_1 et c_2 le plus éloigné de la racine. Ainsi, la mesure de Wu et

Palmer (1994) :

$$Sim(c_1, c_2) = \frac{2 \times prof(c)}{(prof(c_1) + prof(c_2))} \tag{2.3}$$

Avec prof(ci) pour la profondeur du concept ci, c'est-à-dire la distance à la racine de ci ; et c le plus petit ancêtre commun à c_1 et c_2. Certaines autres prennent en compte la profondeur de la hiérarchie, comme avec Leacock et Chodorow (1998), ou encore le type de relation entre les concepts (Hirst et al., 1998). Tout à fait différemment, des approches « basées sur les nœuds », cherchent le contenu informatif des nœuds. Deux versions d'approches existent, la première utilise un corpus d'apprentissage (Chanier & Ciekanski, 2010) et mesure la probabilité de trouver un concept ou un de ses descendants dans ce corpus. Soit c un concept, et p(c) la probabilité de trouver un de ses descendants dans le corpus. Le contenu informatif associé par :

$$IC(c) = log(p(c)). \tag{2.4}$$

Si nous cherchons la similarité entre les concepts c_1 et c_2, il nous faut alors trouver l'ensemble des concepts qui les subsument tous les deux. Soit S (c_1, c_2) cet ensemble. Selon Resnik (1995), nous avons par exemple :

$$Sim_{Resnik(c_1,c_2)} = maxc \in S(c_1, c_2), [IC(c)] \tag{2.5}$$

La seconde version refuse l'utilisation d'un corpus et essaie de calculer le contenu informatif des nœuds à partir de WordNet de Miller (1998) uniquement. La thèse de Seco et al. (2004) est un enrichissement sémantique de requêtes utilisant un ordre sur les concepts. Plus un concept a de descendants, moins il est informatif. Ils utilisent donc les hyponymes [21] des concepts pour calculer le contenu informatif de ceux-ci.

$$ic_{un(c)} = \frac{\frac{log((hypo(c)+1)}{max_{un})}}{log(\frac{1}{max_{un}})} = \frac{1 - log(hypo(c) + 1)}{log(max_{un})} \tag{2.6}$$

Avec hypo(c) qui indique le nombre d'hyponymes dont dispose le concept c, et maxun, une constante qui indique le nombre de concepts, les différentes mesures de simila-

21. Terme dont le sens est compris dans celui d'un autre plus général.

rité sémantique utilisant le contenu informationnel de Resnik (1995) peuvent donc être redéfinies en utilisant celui de Seco et al. (2004). Les deux grandes approches définies précédemment peuvent être combinées. Souvent, il s'agit de réutiliser le contenu informatif et le plus petit ancêtre commun (c), comme avec Lin et al. (1998) :

$$Sim_{lin}(c_1, c_2) = \frac{2 \times log(P(c)}{log(P(c_1) + log(P(c_2)} \tag{2.7}$$

Bidault (2002) propose une numérotation de tous les concepts de l'ontologie, en partant du principe que descendre, se spécialiser, c'est acquérir des caractéristiques. Ainsi, en regardant le ou les numéros d'un concept, on peut facilement savoir non seulement quelle est sa profondeur, mais aussi quels sont ses ancêtres, leur nombre, etc. Nous présentons des formules quelque peu modifiées par rapport à celles de Bidault (2002), car les siennes ne sont pas « normalisées » et ne permettent pas de « ventiler » les concepts sur tout l'intervalle des valeurs de similarité. Soient deux descripteurs m_j et n_i, nous avons la note de proximité de m_j centré sur ni :

$$R_{(m_j \longrightarrow n_i)} = \frac{2^{P_h - P_{com_{ij}+1}} - 2^{P_h - P_{n_i+1}}}{P_h} - M \times (|m_j| - |com_{ij}|) \tag{2.8}$$

Avec com_{ij} la partie commune aux deux descripteurs, $Pcom_{ij}$ qui est la profondeur du descripteur commun à ni et m_j, Ph la profondeur de la hiérarchie, Pni la profondeur d'un descripteur et M, un malus. Selon nous, le malus vaut 1/(Ph)2 pour permettre de « ventiler » tous les descripteurs selon leur proximité au descripteur pivot, c'est-à-dire les répartir sur tout l'intervalle de valeurs. Nous avons ensuite les fonctions permettant de noter la proximité d'un concept c centré sur un descripteur ni, puis d'un concept c centré sur un autre c' :

$$R_{(c \longrightarrow n_i)} = max(R_{(m_j^p \longrightarrow n_i)}, p \in [1....k]) \tag{2.9}$$

$$R_{(c \longrightarrow c')} = moy(R_{(c \longrightarrow n_i^p)}, p \in [1....k]) \tag{2.10}$$

Pour procéder à la formation des communautés, nous avons implémenté dans cette thèse, un algorithme de calcul de similarité sémantique en utilisant les mesures de Lin

et al. (1998) et de Resnik (1995), présentée ci-dessus. En plus de ces mesures, notre algorithme nous permet de créer un modèle d'apprentissage automatique supervisé pour chaque communauté. Le fait de cumuler les technologies du web sémantique et celles de l'apprentissage automatique permet de renforcer le traitement de ce processus de formation et d'évaluer le résultat du calcul de la similarité sémantique.

Les mesures de similarité sémantique présentées précédemment ont été définies sur la base du calcul de l'information (CI) partagée par les concepts concernés. Voici quelques critères de bases que nous avons utiliser dans cette thèse :

- CI du pppc des concepts c_1 et c_2 : CI[pppc(c_1,c_2)] ;
- CI des concepts c_1 et c_2 qui sont compares : CI(c) (c est un concept) ;
- utilisation de ressources externes (corpus) : RE ;
- évaluation du nombre de descendant d'un concept c : hypo(c).

Nous nous baserons sur le tableau 2.2 pour effectuer nos analyses.

Mesures	Année	CI[pppc(c_1,c_2)]	CI(c)	RE	hypo(c)
Resnik	1995	oui	non	oui	non
Lin	1998	oui	oui	oui	non
Resnik(CI)	2004	oui	non	non	oui
Lin(CI)	2004	oui	oui	non	oui

TABLE 2.2 – Comparaison de mesures de similarités basées sur l'CI

*

* *

Dans ce chapitre, nous avons abordé dans une première partie, la démarche de l'expertise humaine pour guider l'humain dans ses choix. Ensuite, en deuxième partie nous avons décrit les outils sémantiques permettant de faire une représentation informatique des connaissances de l'expertise humaine.

Une présentation de travaux en sciences cognitives a été réalisée notamment, pour distinguer les différents systèmes de représentation des connaissances qui prennent en compte la forme de représentation schématisée. Nous avons aussi vu que la notion de modélisation

des connaissances conduit à l'élaboration de la représentation sémantique d'un domaine. En d'autres termes, la représentation obtenue en image du modèle mental des connaissances est indépendante du formalisme de représentation de ces connaissances. Il est possible par exemple de générer une phrase en roumain et une phrase en français d'une même représentation sémantique.

On peut souligner qu'il est concevable de transposer, au niveau organisationnel, le modèle de gestion des connaissances utilisé au niveau individuel au niveau organisationnel. En effet, les organisations ont appris à gérer de multiples informations, notamment avec l'aide de l'informatique. Obtenues de façon brute, l'ensemble des données de l'organisation se retrouve de manière désorganisé, sans lien, sans référence dans les multiples ordinateurs de l'organisation. Afin de donner un sens à l'information, l'organisation doit se doter d'un système de représentation sémantique et caractériser les données. Nous avons vu dans ce chapitre que c'est le rôle principal de l'ontologie et des métadonnées. Ces structures d'informations offrent le support matériel nécessaire à la réalisation des processus de traitement de la connaissance, qui sont la création, le référencement sémantique, la recherche et l'utilisation de l'information.

Dans le chapitre suivant, nous allons aborder l'une des contributions de cette thèse, la modélisation de notre méthode hybride pour une recommandation sémantique enrichie.

Chapitre 3

Méthode hybride pour un système de recommandation sémantique enrichi

Sommaire

« Tout commence, toujours, par une innovation,
un nouveau message déviant, marginal, modeste,
souvent invisible aux contemporains. »
Edgar Morin

En début de thèse, le logiciel Retaily se limitait au recueil de données issues d'enquêtes mystères et à la génération de rapports. Ces rapports contiennent des graphes statistiques et ne permettent pas de faire des recommandations aux commanditaires des enquêtes mystères. Le seul moyen d'obtenir des recommandations est de faire intervenir un expert humain, qui doit analyser attentivement les rapports produits par Retaily, et une augmentation importante des coûts pour les commanditaires des enquêtes mystères. En effet, chaque rapport contient de nombreuses données hétérogènes (graphiques, chiffres, historique d'anciennes recommandations, etc.), qui exigent un temps d'analyse important de la part de l'expert, en charge également de la formulation et de la rédaction des recommandations personnalisées pour chaque point de vente.

L'objectif de cette thèse est d'automatiser partiellement le processus d'analyse des résultats d'enquête mystère et de proposer à l'expert une liste de recommandations, dont il estimera la pertinence. Pour viser cet objectif, nous avons identifié plusieurs défis à résoudre :

— l'analyse fastidieuse des rapports d'enquêtes mystères par l'expert ;
— l'hétérogénéité des données issues des enquêtes mystères ;
— le coût important pour trouver les axes à améliorer au sein de chaque point de vente ;
— la pertinence des recommandations que l'expert va exprimer.

Les travaux scientifiques que nous avons sélectionnés dans les deux premiers chapitres ne nous permettent pas de résoudre directement ces défis.

Ainsi nous avons conçu et développé un système de recommandation sémantique enrichi (SRSE). Dans ce chapitre, nous décrivons la conception de notre SRSE. Nous allons présenter dans un premier temps le SRSE, ensuite dans un second temps, les différents modules qui le compose et leurs différents composants.

3.1 Système de recommandation sémantique enrichi (SRSE)

Notre méthode de conception du SRSE intègre les technologies du web sémantique, de l'apprentissage automatique et la méthode de filtrage collaboratif (FColl).

3.1.1 Pourquoi SRSE ?

Pour améliorer le processus d'analyse de l'expertise humaine, notre SRSE offre les apports suivants : l'automatisation partielle de la démarche d'analyse, l'homogénéisation des données, l'amélioration de la prédiction pour la recommandation et l'amélioration du démarrage à froid pour un nouveau point de vente.

L'automatisation partielle de la démarche d'analyse est l'apport principal de cette thèse. Nous proposons d'automatiser les méthodes utilisées par l'expert pour faire l'analyse des données issues des enquêtes mystères dans le but de formuler des recommandations aux différents points de vente. Cette automatisation permet de réduire son temps de travail et le coût pour les commanditaires des enquêtes mystères.

L'homogénéisation des données va nous permettre d'obtenir une interopérabilité sur l'ensemble des données intervenant dans le traitement de la recommandation, mais aussi pour que le sens de ces dernières soit décrit de manière compréhensible tant par les hommes que par les machines (Berners-Lee et al., 2001). En effet, les résultats produits par le système Retaily sont très hétérogènes (graphiques, chiffres, historique d'anciennes recommandations, etc) et sont associés à des domaines divers. Notre méthode d'homogénéisation est basée sur les technologies du web sémantique.

Pour un système de recommandation (SR), la prédiction est un traitement permettant d'anticiper la valorisation ou la préférence qu'un utilisateur attribuerait par exemple à un produit (Blandin et al., 2019). Dans cette thèse, nous avons décidé d'améliorer ce traitement de prédiction des SRs traditionnels en utilisant les technologies de l'apprentissage automatique. Nous proposons ainsi, une méthode de recommandation basée sur le FColl comportemental de (Esslimani & Igalens, 2008) et sur les algorithmes de l'apprentissage automatique.

Comme vu dans le chapitre 1 page 50, le problème du démarrage à froid entraîne de très mauvais résultats pour les nouveaux points de vente. Ce problème de démarrage à froid intervient lorsque le système accueille un nouveau point de vente, le profil de ce dernier existe, mais ne dispose pas suffisamment de données. Cela conduit à des recommandations non adaptées au profil du nouveau point de vente (voir chapitre 1, page 50). Nous proposons une méthode qui améliore le démarrage à froid en utilisant les notions de similarité sémantique entre les profils des points de vente abordés dans le chapitre 2 page 85.

3.1.2 Architecture global SRSE

L'architecture de SRSE se compose de trois modules consacrés à (voir Figure 3.1) :

— la préparation et à la représentation des données ;

— la prédiction de recommandations ;

— la classification des items à recommander.

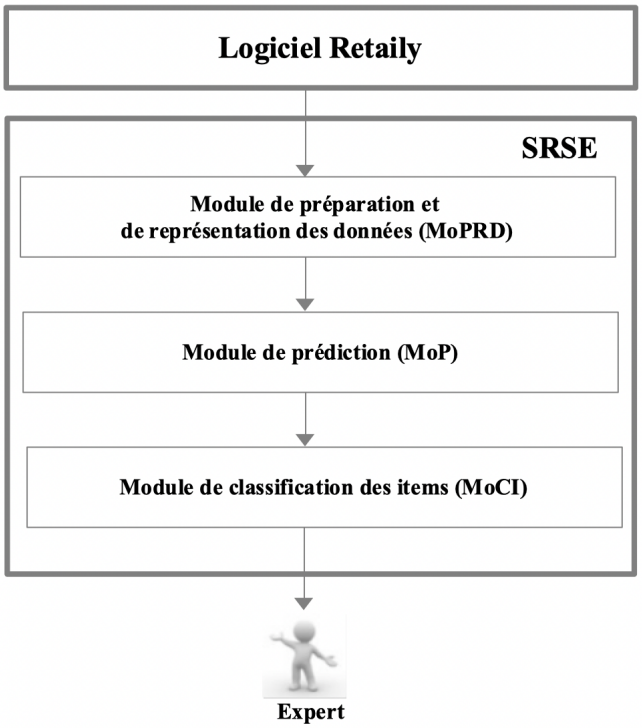

FIGURE 3.1 – Architecture du SRSE

Le module de préparation et de représentation des données (MoPRD) permet de préparer et de faire la représentation des données pour le traitement des recommandations. Ce module est basé sur les technologies du web sémantique et sur des méthodes de l'apprentissage automatique. Le MoPRD reçoit en entrée les données des résultats d'enquêtes mystères produits par Retaily.
La préparation des données consiste à sélectionner, à anonymiser et à décrire les résultats issus des enquêtes mystères, les données des points de vente et les historiques de recommandations. Cette préparation nous permet aussi d'extraire des connaissances à partir des données sélectionnées et d'en faire une représentation. Cette représentation des connaissances est effectuée par les outils du web sémantique (WS). La sortie obtenue de ce module est un ensemble de matrices sous format Json.

Le module de prédiction (MoP) traite les matrices pour produire des prédictions et dégager une liste primaire d'items à recommander. Il est basé sur la méthode du FColl et sur les technologies de l'apprentissage automatique. Le MoP reçoit en entrée l'ensemble des matrices produit par le MoPRD précédent. Le traitement qu'il réalise décrit à travers un algorithme prédictif qui repose sur l'apprentissage automatique non supervisé.

Le module de classification des items (MoCI) classifie les items de liste primaire pour guider l'expert dans sa prise de décisions définitives afin de les présenter aux points de vente. En effet ce module nous permet de classer les items de la liste primaire obtenue du MoP en fonction de la pertinence de ces derniers pour l'amélioration des forces de ventes des différents points de vente. Cette classification est basée sur un algorithme d'apprentissage automatique supervisé. L'algorithme va générer en sortie un fichier exploitable par l'expert au format PDF comportant la liste des items à recommander classée en fonction de leur pertinence.

3.2 Module de préparation et de représentation des données (MoPRD)

Dans le MoPRD nous procédons au traitement des données issues des enquêtes mystères réalisées par le biais de la plate-forme Retaily. Par exemple, pour chaque point de vente, la plate-forme Retaily collecte la géolocalisation, l'organisation des produits, le taux d'affluence par tranche horaire, etc. À ces données s'ajoutent les données d'anciennes enquêtes mystères, ainsi que les recommandations produites pour chaque point de vente. Dans ce module l'hétérogénéité des données est l'un des verrous majeurs à lever. Pour la conception de ce module, nous nous sommes appuyés sur les technologies d'apprentissage automatique non supervisé et celles du Web Sémantique (WS), permettant une représentation des connaissances et des inférences automatiques.
L'architecture de notre module s'articule en quatre composants (voir Figure 3.2). Les données sont d'abord traitées par le composant de sélection de données (COSD). Puis, par le composant d'extraction des connaissances (CoEC). Ensuite, le composant de représentation des connaissances et homogénéisation des données (CoRCHD). Enfin, les données enrichies sont traitées par un composant de formation de communautés de points de vente et de création des matrices (CoFCPV).

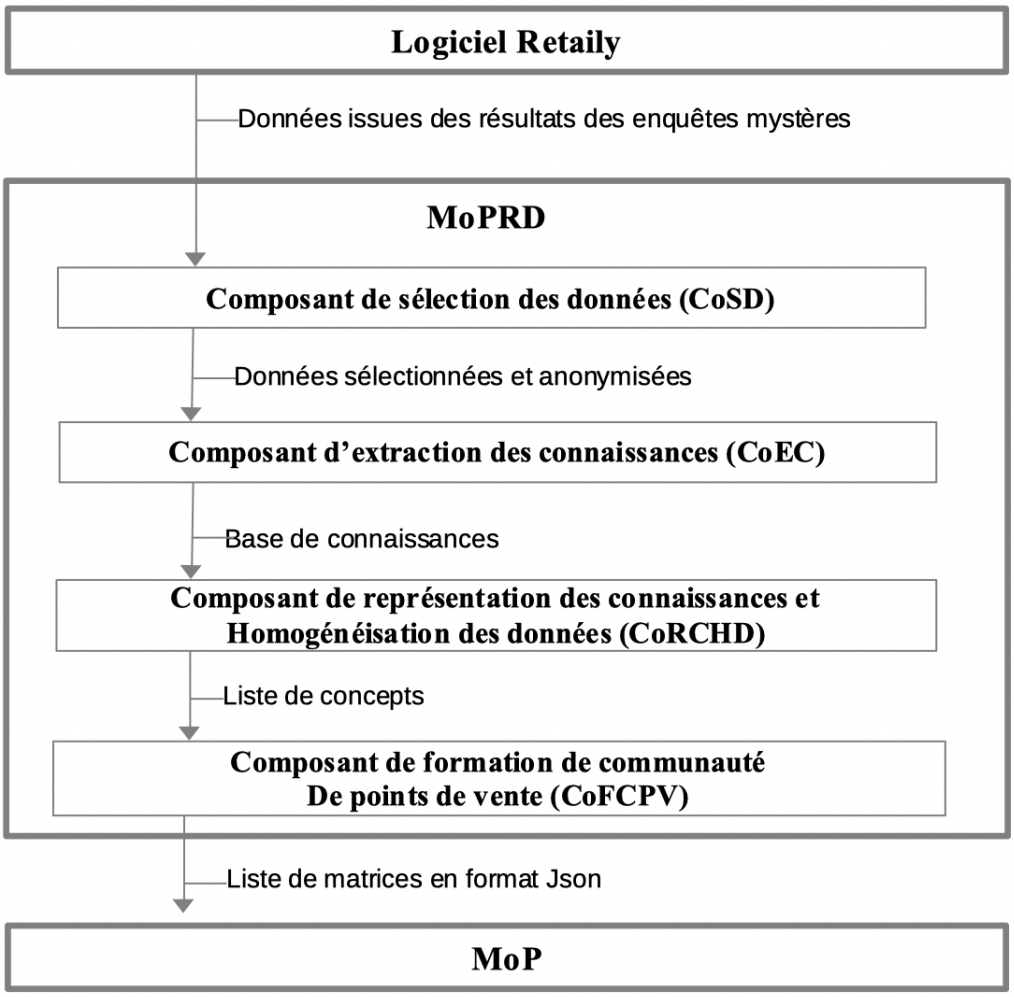

FIGURE 3.2 – Architecture du MoPRD

Le composant de sélection de données (CoSD) est le premier composant du MoPRD et reçoit en entrée les données des résultats issus des enquêtes mystères. Le traitement qui est réalisé dans le CoSD est la sélection des données parmi celles qu'il a reçues en entrée. La sélection est réalisée sur la base d'un ensemble de critères fixés par les commanditaires des enquêtes mystères. En plus de la sélection des données, l'anonymisation des données sélectionnées est effectuée. Cette anonymisation a pour but de respecter les normes qui régit du règlement général sur la protection des données (RGPD). La sortie obtenue du CoSD est l'ensemble des données sélectionnées et anonymisées décrit avec le langage JSON.

Le composant d'extraction de connaissances (CoEC) est le deuxième composant du MoPRD, il nous permet d'extraire des connaissances à partir des données qui ont été sélectionnées et anonymisées dans le CoSD. Nous avons décidé de réaliser l'extraction des connaissances à partir des données pour préparer et analyser les données dans le but de les transformer en connaissances et pour ensuite faire une représentation des connaissances. Pour réaliser l'extraction des connaissances à partir des données nous avons exploité les algorithmes relevant du domaine de la fouille de données. Le CoEC produit ainsi en sortie une base de connaissances décrite avec le langage Json.

Le composant de représentation des connaissances et d'homogénéisation des données (CoRCHD) succède au CoEC et permet de faire la représentation des connaissances et d'homogénéiser les données. Le CoRCHD est utilisé dans notre MoPRD pour lever les verrous de l'hétérogénéité des données. En effet, la représentation des connaissances va nous permettre de structurer et de décrire les données, de les référencer et aussi de définir les relations qu'elles ont entre elles. Pour résoudre cette problématique d'hétérogénéité, nous avons exploité les avantages des ontologies pour une modélisation conceptuelle partagée et partielle de la base de connaissances obtenue dans le CoEC.

Le composant de formation de communautés de points de vente (CoFCPV) reçoit en entrée l'ensemble des concepts issus de la représentation sémantique. La formation des communautés de points de vente (chapitre 1 page 12-13) est réalisée à l'aide d'une métrique calculant la similarité sémantique des différents profils de ces points de vente. Dans le CoFCPV, l'apprentissage automatique non supervisé est utilisé pour enrichir le traitement à la formation de ces communautés.

À partir des communautés de points de vente, nous construisons l'ensemble des matrices nécessaires pour le traitement de la recommandation.

3.2.1 Composant de sélection des données (CoSD)

La Figure 3.3 ci-dessous illustre le processus de sélection et d'anonymisation des données..

Afin de sélectionner les données nous avons formulé et implémenté plusieurs requêtes Sql sous MySql (voir dans l'annexe B).

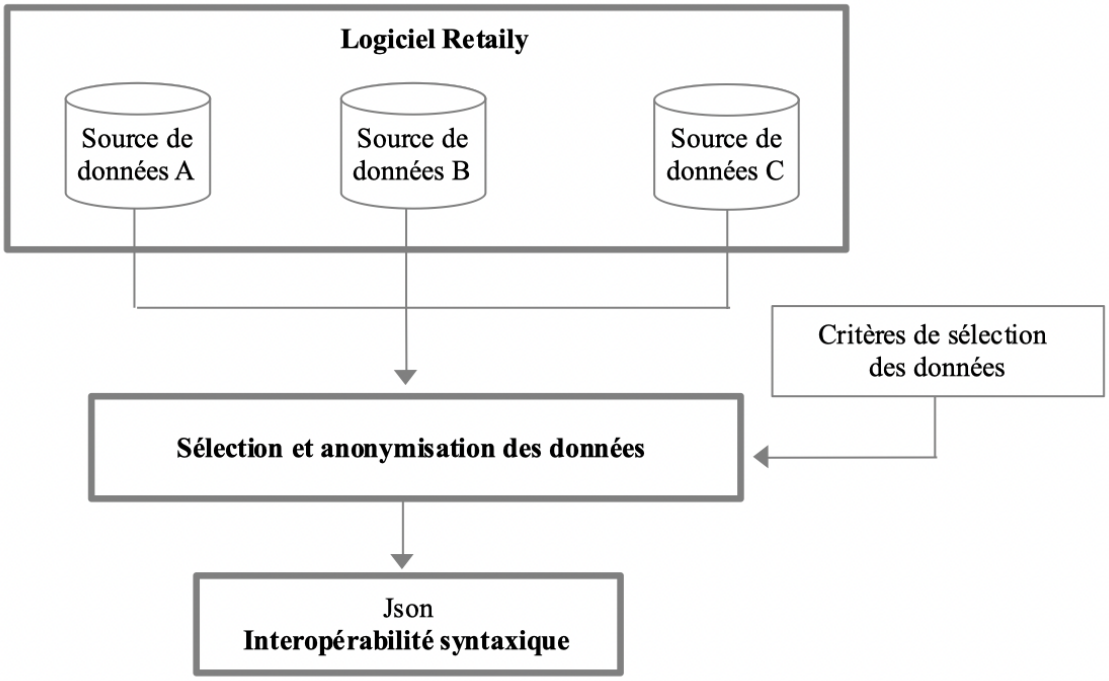

FIGURE 3.3 – Architecture du CoSD

La sélection des données est réalisée sur la base d'un ensemble de critères de sélection des données qui est défini par les commanditaires des enquêtes mystères sur la plate-forme Retaily. Ces critères de sélection correspondent aux différents axes que les commanditaires souhaitent évaluer à travers les enquêtes mystères.

Concernant l'anonymisation, nous avons développé un script en PHP pour parcourir les données et de les renommer selon le cadre juridique imposé par la RGPD ou la CNIL.

Les résultats de sélection et d'anonymisation sont décrits à l'aide du langage Json permettant d'optimiser les requêtes et la récupération de données.

Le langage Json permet l'interopérabilité syntaxique au niveau SRSE, c'est-à-dire que les données sont structurées selon un format commun dans tout le système facilitant la

réutilisation des données et le partage de ces dernières entre les différents modules ou composants du SRSE.
Avec cette structuration des données sélectionnées et anonymisées, le traitement de notre SRSE se poursuit en faisant appel au composant d'extraction des connaissances.

3.2.2 Composant d'extraction des connaissances à partir des données (CoECD)

Le CoECD traite non seulement les données sélectionnées et anonymisées par le CoSD, mais aussi les données relatives aux recommandations reçues dans le passé pour chaque point de vente et les évaluations réalisées à l'aide de la plate-forme Retaily.
Pour réaliser l'analyse des données et en extraire des connaissances, nous allons utiliser des algorithmes relevant du domaine de la fouille de données (FD), présentées dans le chapitre 2.
Le processus d'extraction des connaissances à partir des données (ECD) est divisé en deux phases comme le montre la figure 3.4 : une première phase de prétraitement et une deuxième phase de fouille de données (Fayyad et al., 1996).

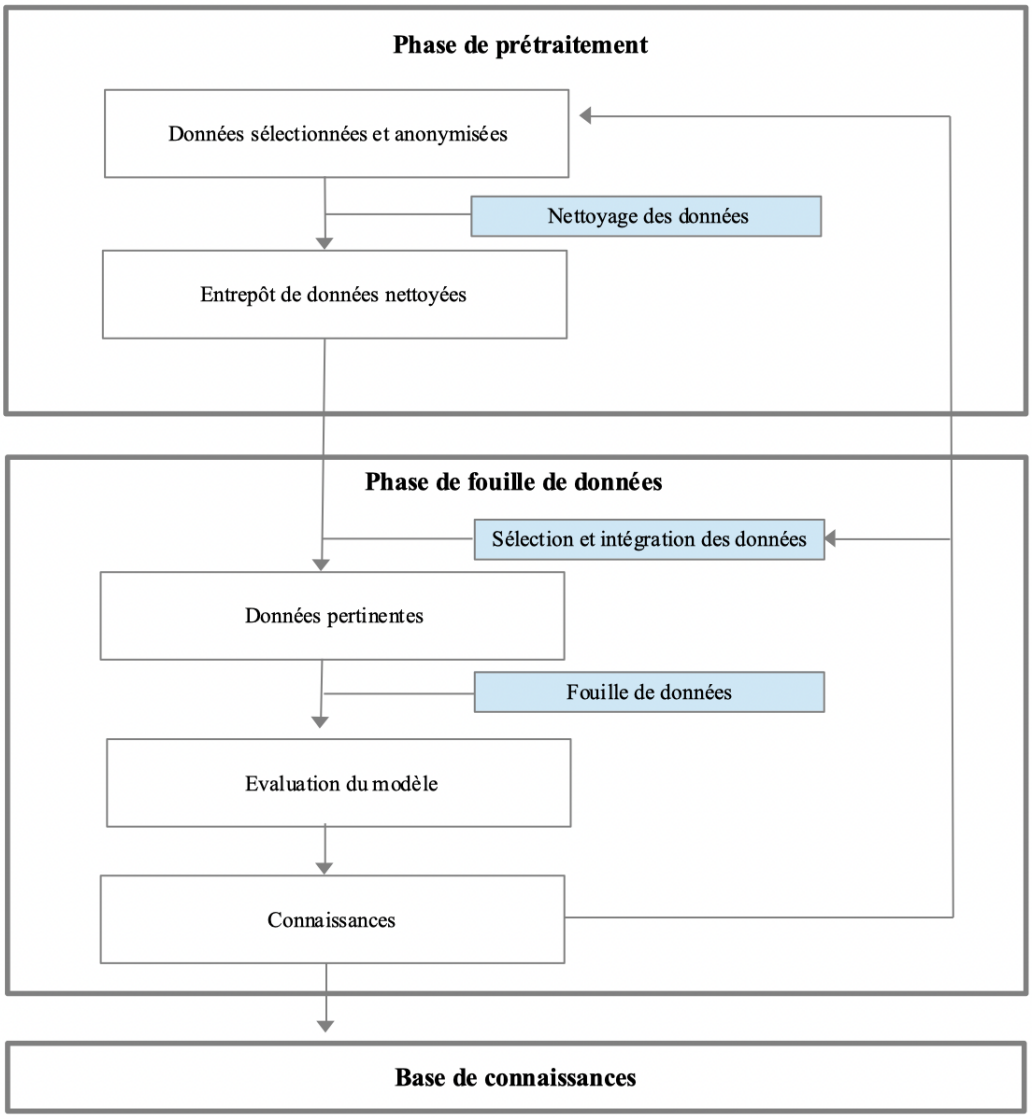

FIGURE 3.4 – Processus d'ECD

Le processus d'ECD commence par une première phase de prétraitement dans l'objectif de supprimer le bruit (données non utiles pour l'ECD, par exemple une fausse adresse postale d'un point de vente), pour traiter les données manquantes et obtenir des données pertinentes. Ainsi, le prétraitement procède à un nettoyage des données sélectionnées et anonymisées dans le CoSD pour former un entrepôt de données nettoyées.
Ensuite, ces données nettoyées vont être sélectionnées puis intégrer dans une deuxième phase pour définir les données pertinentes par rapport aux objectifs de la fouille de données. À la suite de la définition des données pertinentes, le traitement pour la fouille de données est lancé, afin de classifier, de chercher des modèles et de définir des pa-

ramètres appropriés. Ensuite pour déduire les connaissances qui vont être stockées dans la base de connaissances, une évaluation est réalisée. Cette évaluation est faite sur les modèles et qui ont été trouvés dans le traitement de la fouille des données.
Dans le chapitre 2 page 73, nous avons vu que la fouille de données propose deux méthodes : méthodes descriptives et prédictives. Dans cette thèse, nous nous sommes intéressés à la méthode prédictive, car cela permet de générer une première prédiction sur les données pertinentes pour construire notre base de connaissance. La méthode prédictive se compose de deux algorithmes : un premier algorithme de classification supervisée et un deuxième algorithme de régression. Nous avons opté pour l'utilisation de l'algorithme classification supervisée car nous facilitant par la suite à la formation des communautés de point de vente que nous allons aborder dans la section 4.
Le résultat obtenu est un ensemble de connaissances décrites sous le format Json. Le choix de ce format est motivé par la suite du traitement dans le composant suivant qui nous permet de faire la représentation des connaissances à partir des données.
Une fois l'extraction des connaissances effectuée, nous passons à la représentation de ces connaissances pour une homogénéisation des données qui ont été sélectionnées.

3.2.3 Composant de représentation des connaissances et d'homogénéisation des données (CoRCHD)

La méthode utilisée dans le CoRCHD nous permet de faire une représentation structurée des connaissances en utilisant les outils sémantiques, notamment le RDF et l'ontologie. La représentation des connaissances offre la possibilité d'utiliser une ou plusieurs ontologies qui sont spécifiques à un domaine précis (voir chapitre 2, page 64). Les ontologies que nous avons utilisées dans nos expérimentations permettent de conceptualiser les données qui ont été définies pour la recommandation.
Il est important de souligner que des divergences sur la formulation d'une ontologie ont donné naissance à deux méthodes de composition ontologique en fonction du formalisme des relations qu'entretiennent les concepts entre eux (Kassel, 2018).
La première méthode s'oriente vers la composition des ontologies qualifiées d'informelles. Les ontologies informelles ne se limitent pas à des relations de spécifications (El Bouhissi

et al., 2020). Ces dernières regroupent toutes les relations possibles entre les différents concepts représentés dans la structuration sémantique. De ce fait, on retrouve dans la revue de littérature beaucoup de recherches sur l'ontologie informelle ou encore appelée réseau sémantique(Desclés, 1987 ; Chantrain, 2017).

La deuxième méthode permet d'organiser des concepts sur la base d'une hiérarchisation, utilisant les relations de spécifications de type « est un ou est une sorte » de Sowa (2000). Ainsi, l'ontologie est vue comme une approche de sémantisation des données pour un domaine précis. Nous allons-nous orienter vers les ontologies utilisant la deuxième méthode, car permettant de décrire des concepts et des sous-concepts et donnant la possibilité de vérifier la cohérence, la complétude ou encore la non-redondance (voir chapitre 2, page 78). Cette démarche de sémantisation des données nous conduit à une homogénéisation des données sélectionnées et anonymisées.

Ainsi, à partir de la représentation des connaissances, l'ensemble des concepts est recensé pour la construction matricielle sous la forme d'association (point de vente, note, concept). La note représente dans l'association le score ou l'évaluation que le point de vente a obtenue dans le passé sur le concept (voir chapitre 1, page 15).

Après l'homogénéisation des données, nous passons à la prédiction pour la recommandation (Blandin et al., 2019). Pour proposer des recommandations pertinentes, l'étape de la prédiction est incontournable, car permettant d'anticiper sur l'appréciation des items.

3.2.4 Composant de formation des communautés de points de vente (CoFCPV)

Le CoFCPV nous permet de réaliser la formation des communautés de points de vente qui consiste à regrouper les points de vente sur la base de leur similarité sémantique (Slimani, 2013). Cette similarité sémantique est une mesure permettant d'identifier dans un ensemble de points de vente les profils similaires, c'est à dire possédant des concepts, des critères ou encore un historique de recommandations similaires.

Le processus de formation de communautés va réduire le temps de calcul de la similarité sémantique entre les profils des points de vente. En effet, dés qu'un point de vente P est affecté à une communauté, le système n'aura plus à recalculer sa similarité avec les

autres profils des points de vente de sa communauté tant que son profil ou son historique n'auront pas été mis à jour.

Ainsi, avec la formation de communautés un premier filtre est réalisé sur les items susceptibles d'intéresser le point de vente ayant sollicité la recommandation ; cela réduit le périmètre de traitement à la communauté qui est concernée par la recommandation. La Figure 3.5 ci-dessous est une illustration de la formation de communautés de points de vente.

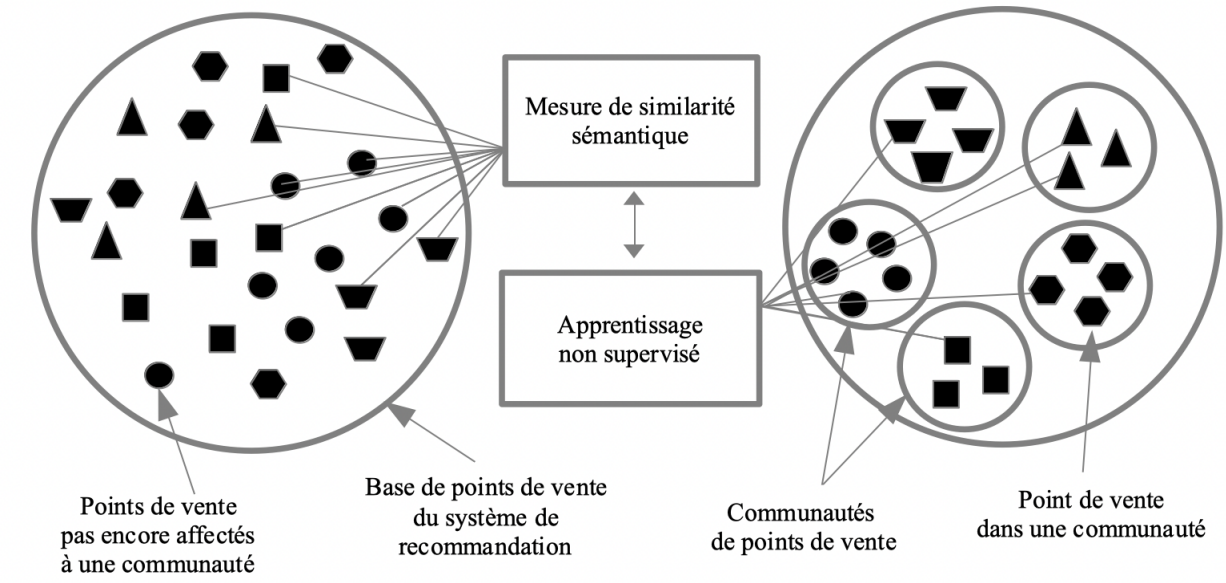

FIGURE 3.5 – Composant de formation de communautés de points de vente

Dans le chapitre 2 page 87, nous avons vu qu'il existe de nombreuses mesures de similarité sémantique, avec des propriétés et des résultats différents et une comparaison de ces mesures de similarité sémantique a été réalisée. Dans cette thèse, nous souhaitons classer tous les concepts par rapport à un concept central, c'est-à-dire un concept de référence ou encore la racine. Il s'agit donc pour nous de choisir la meilleure mesure de similarité sémantique, étant données ces contraintes et eu égard aux résultats des différentes mesures de similarité. Nous sommes dans le cadre d'un graphe dont les nœuds sont des concepts. Il paraît donc évident d'utiliser les chemins (suite d'arcs du graphe) pour mesurer la distance entre les concepts. Dans cette thèse nous avons choisi de travailler avec les approches proposées par Rada et al. (1989) ; Resnik (1995) car, leurs travaux s'orientent vers l'utilisation des ontologies possédant une organisation des concepts sur

la base d'une hiérarchisation.

Pour procéder à la formation des communautés, nous avons implémenté un algorithme de calcul de similarité sémantique en utilisant les mesures de Lin et al. (1998) ; Resnik (1995), présentée dans le chapitre 2 à la page 88-90. En plus de ces mesures, notre algorithme nous permet de créer un modèle d'apprentissage automatique supervisé pour chaque communauté.

Ainsi notre architecture globale détaillé de MoPRD est illustrée dans la Figure 3.6 :

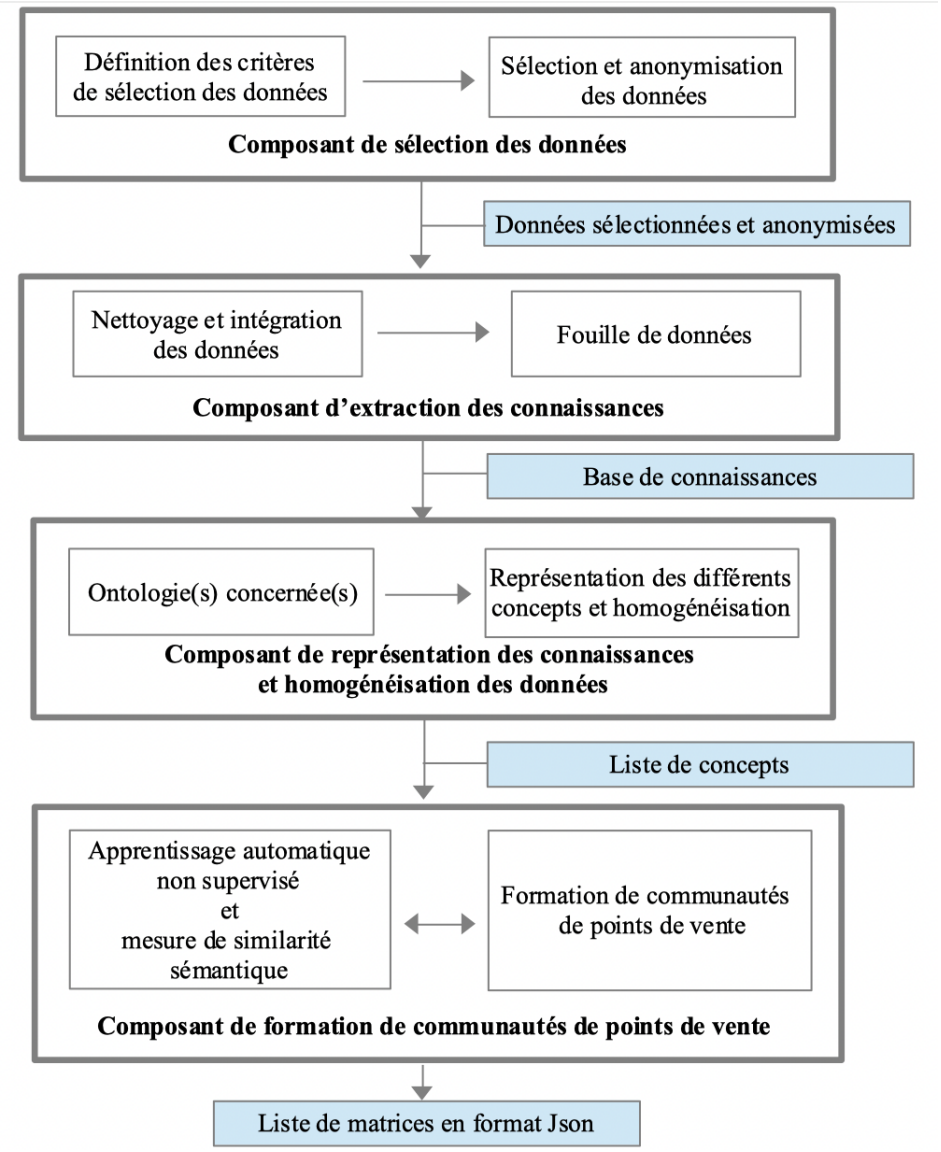

FIGURE 3.6 – Architecture globale détaillé du MoPRD

Au terme du traitement des données par le MoPRD, on obtient une liste de matrices en format Json qui sera reçue en entrée par un module de prédiction. Ainsi, ce module prédiction va succéder le MoPRD dans le traitement des données d'enquêtes mystères par notre SRSE.

3.3 Module de prédiction (MoP)

Le MoP est le deuxième module de notre SRSE, son objectif est d'anticiper l'évaluation qu'un utilisateur d'un SR pourrais donner à un item. Notre MoP est basé sur les technologies de l'apprentissage automatique et sur la méthode du FColl, il est organisé en deux composants : un composant d'apprentissage et un composant de prédiction.

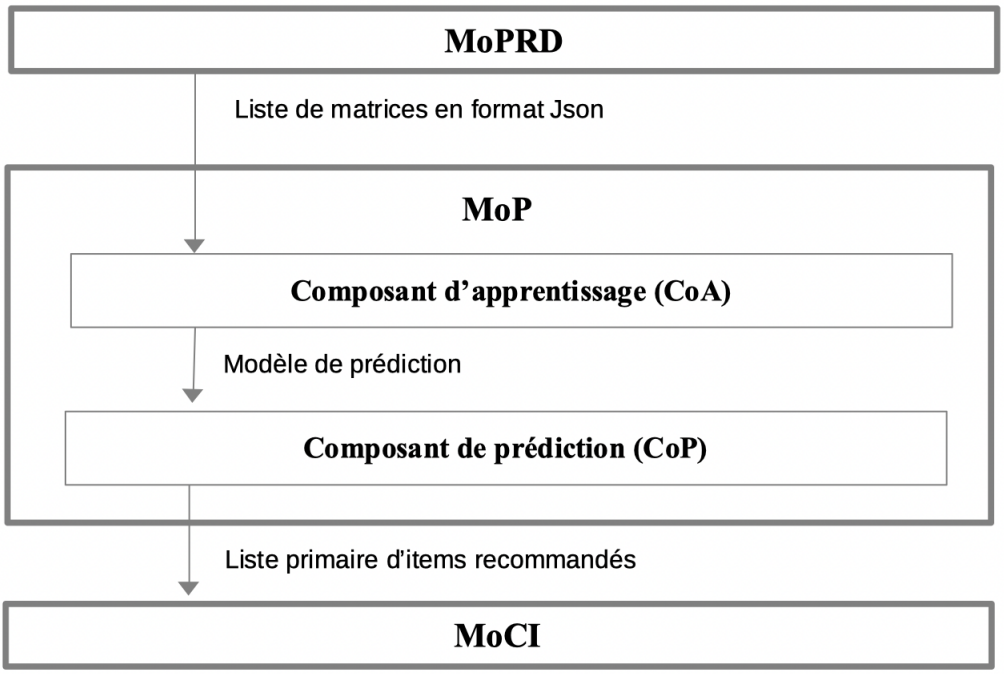

FIGURE 3.7 – Architecture du MoP

Le composant d'apprentissage (CoA) est le premier composant du MoP et a pour objectif d'enrichir le traitement à la recommandation, précisément la prédiction des items à recommander. Il reçoit en entrée l'ensemble des données issues du dernier composant du MoPRD. Le CoA est en charge dans un premier temps de la construction et de la sélection des modèles d'apprentissage (aussi appelé modèle d'entraînement). En

effet le traitement effectué dans ce composant commence par la définition et la sélection de données qui sont les différents concepts issus de la représentation des connaissances effectuée dans le premier module MoPRD de notre architecture globale. Ces données sont sélectionnées sur la base d'un ensemble d'observations. Ensuite, dans un second temps, les données sont découpées et classées soit en données d'entraînement, soit en données de test. Cet ensemble de données d'entraînement ou de test représente les observations permettant de créer nos modèles d'entraînement.

Le composant de prédiction (CoP) succède au CoA, son objectif est de réaliser des prédictions sur la base d'un algorithme prédictif que nous avons implémenté. Le CoP prend le relais en utilisant un algorithme qui se base sur les principes de l'apprentissage automatique non supervisé et du FColl. Ce composant va émettre en sortie un ensemble de prédictions sur les items en fonction des profils de points de vente. Les prédictions vont ensuite être traduites en une liste primaire d'items à recommander. Cette liste va être envoyée au module de classification de notre architecture globale. Le module de classification permet de réaliser la catégorisation des items de la liste primaire sur la base de l'algorithme d'apprentissage supervisé.

3.3.1 Composant d'apprentissage (CoA)

Le CoA est le premier composant du MoP, son objectif est de réaliser le traitement de l'apprentissage automatique supervisé dans le but de construire et d'entraîner un modèle d'apprentissage.

Le composant en charge de la construction et de la sélection du modèle prédictif a pour rôle de construire notre modèle d'apprentissage sur la base d'un ensemble d'observations accessibles (Geer, 2021). Ces observations accessibles sont les données obtenues à partir des toutes les variables du logiciel Retaily. Ces dernières sont exploitées pour trouver d'autres observations dans l'objectif de faire une prédiction sur les valeurs dites explicites. Néanmoins, il reste complexe d'obtenir de bonnes prédictions dans le cas de nouvelles observations.

Pour tout processus de prédiction, la problématique à résoudre est décrite par un ensemble de variables (Gottwald & Reich, 2021). La disposition du nombre d'observations sur les

variables qui sont impliquées est très importante pour la construction d'un modèle à partir des données. Si on prend le cas d'une variable aléatoire, l'observation représente une valeur à l'instant où l'observation a été réalisée. Par l'exemple, dans le cas d'une modélisation composée de plusieurs variables, une observation prend en compte toutes les variables qui sont impliquées au moment de l'observation. Dans le cadre d'une enquête mystère dans un point de vente, les questions qui composent le formulaire d'enquête représentent chacune une variable. Ainsi, pour une partie des réponses du questionnaire saisie à un moment donné, on obtient un ensemble d'observations permettant de créer un modèle partiel d'apprentissage, car ces observations peuvent générer d'autres observations par déduction. Par exemple, si on collecte des réponses aux questions liées à l'accueil des clients au sein du point de vente, on peut générer des réponses pour les questions liées aux compétences du personnel d'accueil du point de vente.

Les données qui permettent de construire ces modèles correspondent à l'ensemble des observations qui incluent une grande partie d'entre elles. Ces données sont des valeurs explicites de la prédiction, c'est-à-dire les variables expliquées. Si on reprend notre exemple, les variables expliquées vont correspondre à l'observation qui a été générée. Il est possible qu'une observation partielle ne permette pas de déduire une autre observation de complétude, car possédant des données clefs manquantes. Ainsi, avant de passer à la sélection du modèle, il est important de procéder à la validation du modèle explicite en s'assurant de ne pas avoir de données manquantes. En effet, un modèle est validé sur la base des observations de même type, c'est-à-dire possédant des variables explicites et celles qui ont servi à la prédiction. La validation le modèle est utilisé seulement pour des observations possédant des variables explicites. Cela permet d'obtenir la prédiction de la variable à trouver.

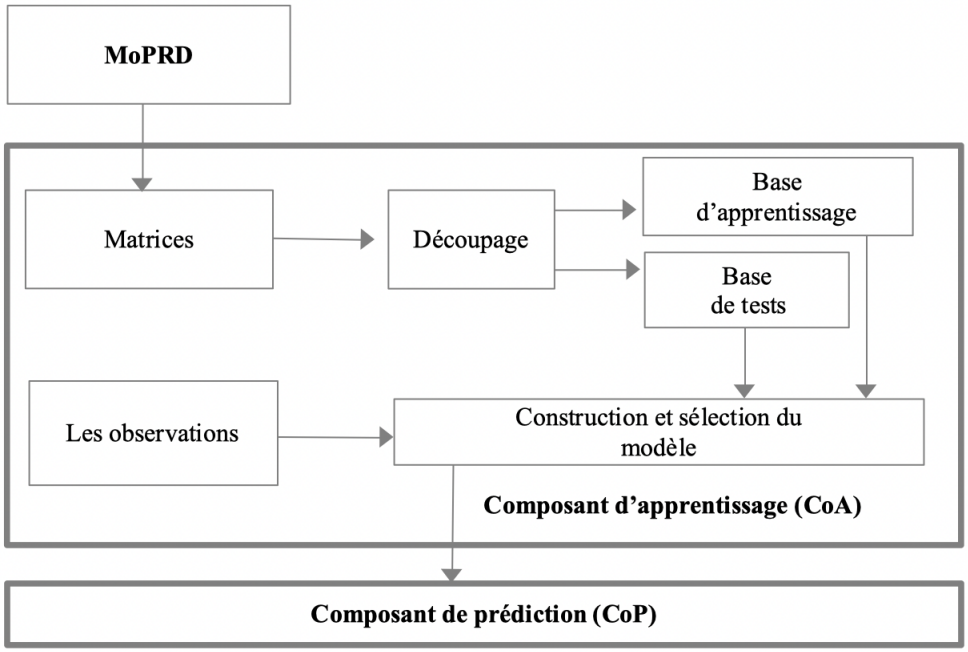

FIGURE 3.8 – Composant d'apprentissage

L'ensemble des données reçues en entrée au niveau du module de la prédiction est découpé en deux sous-ensembles, un sous ensemble pour l'apprentissage et un second pour les tests. Ces sous-ensembles sont appelés respectivement base d'apprentissage (ou d'entraînement) et base de test (Bhavsar & Ganatra, 2012). Ces dernières sont particulièrement formées des données correspondant à l'ensemble des données disponibles. Ces données sont incontournables dans la modélisation de notre processus de recommandation, car il s'agit des données qui sont liées directement au contexte de la recommandation. On peut inclure ensuite d'autres données, dites secondaires qui peuvent compléter l'information contextuelle. Pour procéder au découpage des données et créer notre bas d'entraînement et de test, nous avons utilisé la méthode de la validation croisée.

Validation croisée

Selon Rabut (2020) la validation croisée consiste à découper le jeu de données en plusieurs parties égales. Puis, chacune des parties sont divisées en une base de test et une base d'entraînement. Cette démarche nous a permis d'utiliser l'intégralité des données pour l'entraînement et pour la validation. Il est obligatoire qu'un modèle soit validé après

avoir passé cette étape d'évaluation sur la base des différents tests qui ont été utilisés pour l'apprentissage. Avec l'algorithme des plus proches voisins la liste des erreurs, est toujours vide. En revanche, avec la base de tests, l'efficacité de l'algorithme n'est pas négligeable. Pour confirmer la robustesse de ce dernier, il peut être lancé plusieurs fois selon une technique appelée : la validation croisée. Pour débuter la description de l'ensemble des données, on procède à la construction de modèles utilisateurs, dans notre cas des modèles de points de vente. Cette construction sera basée sur les informations des points de vente et celles des items qui les concernent. Parmi ces informations, on peut citer : les informations propres aux points de vente, l'historique de recherches et d'achats qu'ils ont effectués dans le passé.

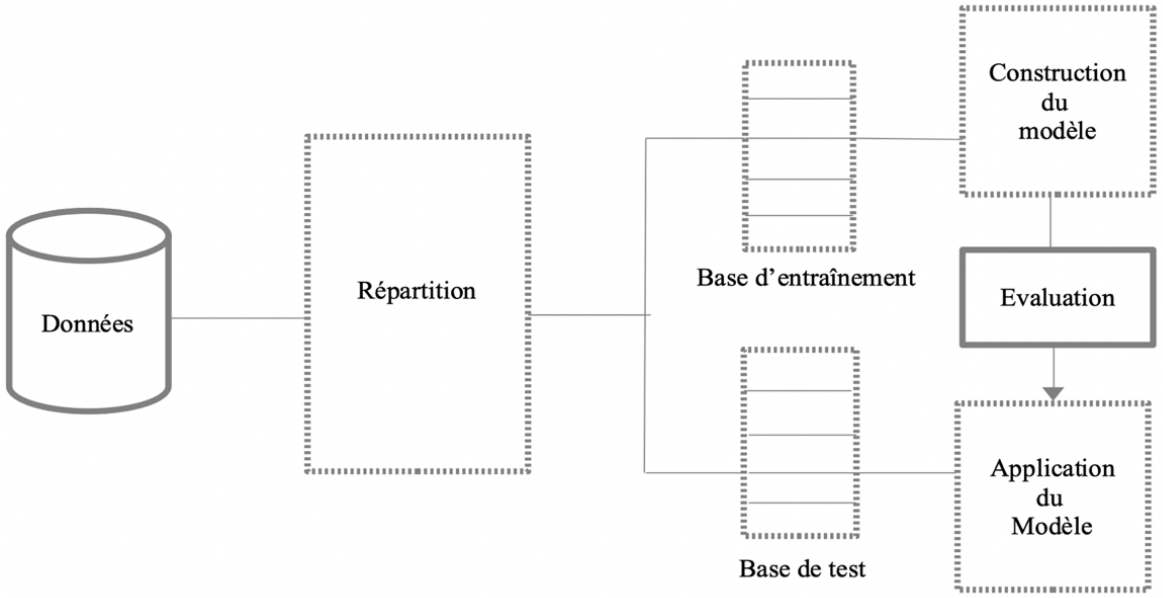

FIGURE 3.9 – Découpage des données par validation croisée

Ces modèles seront utilisés dans le processus de recommandation. Pour ce faire, nous allons employer une technique qui consiste à construire un modèle régressif linéaire qui nous permettra de mesurer la pertinence d'un item i pour un point de vente p. Il s'agit ici de calculer le taux de pertinence pour qu'un item Ii soit intéressant pour un point de vente p. Cela revient à faire une prédiction de l'item i pour le point de vente p, autrement dit évaluer la pertinence de recommandation I_i pour le point de vente p.
Soit L_p une liste de points de vente et I_p, une liste d'item à proposer aux différents points

de vente dans leur démarche d'amélioration.

P_i = la taille point de vente ; I_i= une compétence du personnel et I_p = la localisation point de vente.

$$Prédiction(Pi, Ii) = Pi + Ii + lp \tag{3.1}$$

‘

Le problème de cette approche est le processus de sélection du modèle, par exemple la prise en compte du paramètre date dans la sélection du modèle. En effet, le paramètre « date » doit être en adéquation avec le contexte de la recommandation. Lorsque le nombre de points de vente et d'items est grand, il devient très compliqué d'insérer des variables pertinentes dans le modèle. À titre d'exemple, il est possible qu'un film soit préféré à un autre, construit sur le même scénario, parce qu'il est mieux interprété ou que le scénario a été mieux explicité, donc plus facile à comprendre. Si on se base sur cet exemple on peut se poser la question sur l'aptitude d'un système à évaluer l'interprétation d'un scénario de film. Serait-il nécessaire d'avoir dans ce cas des paramètres de recommandation, des variables décrivant ces facteurs implicite ? Pour apporter des réponses à ces questions, nous avons associé à notre approche de modèle de point de vente, une approche permettant de prendre en compte ces facteurs implicites. Cette nouvelle approche est basée sur la mesure de similarité sémantique et des plus proches voisins (Ferré, 2017). Notre méthode hybride cherche dans un premier temps à former des communautés de points de vente en utilisant les mesures de similarité et des plus proches voisins. Ensuite, elle procède à la construction des modèles pour chaque communauté de points de vente. Ainsi en utilisant les algorithmes de l'apprentissage supervisé, nous pouvons obtenir une classification sur les données qui ont été catégorisées. La technique des plus proches voisins sera utilisée pour les nouveaux points de vente qui ne sont pas encore catégorisés (Guo et al., 2003). De ce fait pour procéder à une validation croisée, nous découpons notre base de données en plusieurs sous ensemble et isolons les éléments les plus recommandés et ceux utilisés dernièrement pour les tests.

Paramètres et hyper-paramètres de l'apprentissage automatique

Les paramètres de l'apprentissage automatique sont l'ensemble des variables utilisées dans l'algorithme d'apprentissage. Ces variables nous permettent de stocker les données destinées au traitement de l'apprentissage automatique. les paramètres de l'apprentissage automatique sont associés à un modèle spécifique d'apprentissage automatique. Ces modèles d'apprentissage automatique sont des algorithmes d'optimisation. Dans un modèle d'apprentissage automatique, il existe 2 types de paramètres :

1. paramètres du modèle : Ce sont les paramètres du modèle qui doivent être déterminés à l'aide de l'ensemble des données de la base d'entraînement ;
2. les hyper-paramètres sont les paramètres réglables qui doivent être ajustés afin d'obtenir un modèle aux performances optimales.

Dans cette thèse, nous nous intéressons aux paramètres de type hyper-paramètres, notamment l'algorithme de descente gradient qui est utilisé dans le cadre de l'optimisation. L'algorithme de descente gradient est itératif et est utilisé dans cette thèse dans le but de réduire les fonctions définies dans notre espace euclidien. Il permet d'obtenir une amélioration de manière successive, cela signifie que sa transition est exprimée par la technique de recherche linéaire tout au long. Il est logique de penser qu'en utilisant des paramètres similaires, on obtienne de meilleurs résultats, quelles que soient les données qui ont été prises en compte. La question que l'on se pose est : comment connaître les paramètres qui permettent d'obtenir de meilleurs résultats ? La technique que nous avons utilisée est la réalisation d'un test sur un ensemble de valeurs et la sélection de celles qui donnent le meilleur résultat. Les limites de l'apprentissage automatique sont non négligeables.

Plusieurs travaux qui ont abordé les technologies d'apprentissage automatique pour la recommandation ont souligné des limites (LeCun et al., 1998 ; Cord & Cunningham, 2008 ; Molnar, 2019). Parmi ces limites, on peut citer :

1. le besoin d'une quantité de données importante pour les tâches complexes. Dans le cas de l'apprentissage supervisé, l'annotation des données est très fastidieuse et prend beaucoup de temps. Son utilisation dans le cadre du traitement des langages naturels est très complexe ;

2. les données d'entraînement peuvent être biaisées ;
3. les problèmes qui sont liés à la classification, à la régression et à la structuration des prédictions. Dans le cas de la classification si on prend deux produits similaires qui sont conçus de manière différente et avec des composants qui ne sont pas les mêmes, alors pour l'algorithme de classification les données brutes comme la taille, le « design » ou encore l'appellation sont évidentes, mais celles qui sont considérées comme des métadonnées sont abandonnées. Il devrait être possible de faire une prédiction sur la base des métadonnées des données qui sont accessibles dans la base. Cette problématique ne concerne pas la régression, car la prédiction n'est pas basée sur une quantité, mais sur une base d'informations.

Au terme de l'entraînement de la base d'apprentissage, on obtient des prédictions et ces dernières sont comparées aux éléments de la base de test pour évaluer la pertinence de la prédiction (Bustillo et al., 2021). La mesure de pertinence d'un modèle de recommandation n'étant pas évidente, une méthode très simple consiste à faire une comparaison entre les items reçus à l'issue du processus de recommandation et les résultats obtenus par l'expert. Le fait que les modèles des plus proches voisins (K-NN) renvoient toujours de bons résultats quand l'item est déjà connu par le système (Bijalwan et al., 2014). Donc il serait pertinent de proposer le même item aux points de vente ayant des profils similaires. Il faudra aussi proposer à ce dernier un nouvel item que l'on supposera être pertinent pour lui ; soit un item similaire à celui qu'il a jugé satisfaisant.

3.3.2 Composant de prédiction (CoP)

Au niveau du composant de prédiction, l'identification des métriques est effectuée pour le processus de raisonnement. Ces métriques représentent les méthodes utilisées par les experts du domaine pour analyser les données d'enquêtes mystères et de réaliser des prédictions (Becker et al., 2010). L'identification et l'application des métriques permettent de réaliser un modèle d'apprentissage et de les intégrer dans notre algorithme de prédiction. Ces métriques sont incontournables pour créer un moteur d'inférence pour la recommandation (Ouellet & Tessier, 1987 ; Tahiraly, 2014). Dans cette thèse, elles sont appliquées sur les concepts qui ont été décrits au niveau de l'ontologie utilisée. L'archi-

tecture illustré dans la Figure 3.10 montre les étapes exécutées dans le CoP.

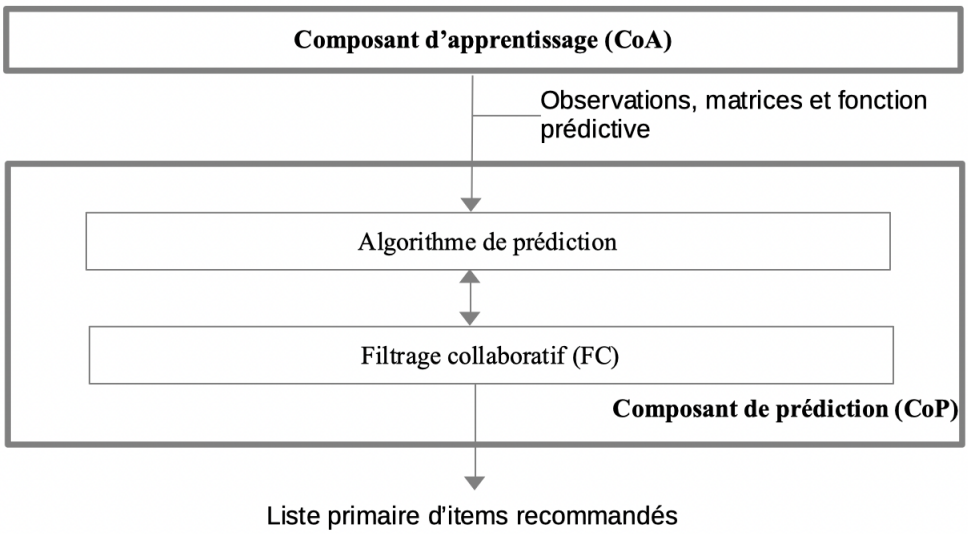

FIGURE 3.10 – Architecture du CoP

Pour implémenter ces métriques, nous avons intégré ces dernières dans notre algorithme de FColl. Cette démarche d'implémentation est basée sur la méthode Delphi (Vernette, 1994) présenté dans le chapitre 2, page 57. Comme nous l'avons vu dans le chapitre précédent, cette méthode est très coûteuse quand elle est réalisée par l'humain, mais très efficace en termes de pertinence des résultats qu'elle propose, mais avec l'utilisation des nouvelles technologies, la problématique du coût est résolue. Ainsi en faisant appel à cette méthode on conserve la pertinence de ses résultats.

Génération de la prédiction et filtrage collaboratif

La génération de la prédiction pour la recommandation est le résultat d'un traitement sur une grande quantité de matrices possédant des items avec des variances faibles par rapport aux points de vente, elle est sous la forme d'appréciations prédites sur ces items. Les appréciations sont obtenues à partir d'un algorithme sur une méthode de prédictions, il est possible d'utiliser plusieurs méthodes différentes. L'une de ces méthodes permet de faire un prétraitement sur les voisins des items avec peu de variabilité afin de réduire les calculs de prédictions très coûteux. Une autre technique est de procéder à une classification non préalable des items ou des points de vente dans l'optique d'optimiser la

« dimensionnalité » de la problématique. Il s'agit d'un réajustement des modèles dans le but d'obtenir une optimisation sur les fonctions qui sont utilisées dans plusieurs dimensions. Cette méthode qui conduit souvent à une dégradation des performances des prédictions. Enfin, les méthodes de réduction de dimensionnalité classique sont également utilisées en prétraitement.

Pour le traitement de la génération des prédictions, nous avons choisi et utilisé l'algorithme K-NN (Ali et al., 2020). L'avantage d'utiliser K-NN est qu'il n'a pas besoin de modèle pour pouvoir effectuer une prédiction. Cependant, son inconvénient est qu'il doit sauvegarder en mémoire toutes les observations pour pouvoir effectuer sa prédiction. De ce fait, il est important de faire attention à la taille de l'ensemble des données de la base d'entraînement (Jadhav & Channe, 2016).

Il est aussi à noter que le choix de la méthode de calcul de la distance ainsi que le nombre de l'algorithme KNN peut ne pas être évident. En effet, Il faut essayer plusieurs combinaisons et faire du « tuning » de l'algorithme pour avoir un résultat satisfaisant (Viola et al., 2019).

Une fois la prédiction effectuée, le filtrage collaboratif est réalisé sur la base des résultats de la prédiction. Ainsi les items ayant reçu les meilleurs résultats de prédiction sont recommandés et inscrits dans une liste primaire dans un but une classification des items qu'on va aborder dans la section IV.

Cet algorithme se base sur le jeu de données en entier. La démarche de ce dernier est la suivante :

- pour une observation ne faisant pas partie du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches de notre observation.
- pour ces K voisins, l'algorithme se base sur leurs variables de sortie y pour calculer la valeur de la variable y de l'observation qu'on souhaite prédire.

On peut décrire le fonctionnement de l'algorithme du K-NN avec le pseudo-code suivant :

Algorithm 1 Notre algorithme de prédiction

Données en entrée :

• l'ensemble des données collectées D ;

• la fonction permettant de définir la distance d ;

Soit X, une nouvelle observation. Pour obtenir la prédiction de X en une sortie y on fait :

• calculer l'ensemble des distances de l'observation X, avec les autres observations l'ensemble des données D ;

• sauvegarder les K observations de l'ensemble des données D les
proches de X en utilisant la fonction de calcul de distance d ;

• Extraire les valeurs de y des D observations sauvegardées :

◦ Si on effectue une régression, calculer la moyenne (ou la médiane) de y retenues

◦ Si on effectue une classification, calculer le mode de y retenues

• Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par K-NN pour l'observation X.

Ce MoP fournit en sortie une liste primaire d'items qui va être utilisée en entrée dans le module suivant, en charge de la classification de cette liste primaire. Ainsi l'architecture globale détaillé de notre MoP est illustré dans la Figure 3.11.

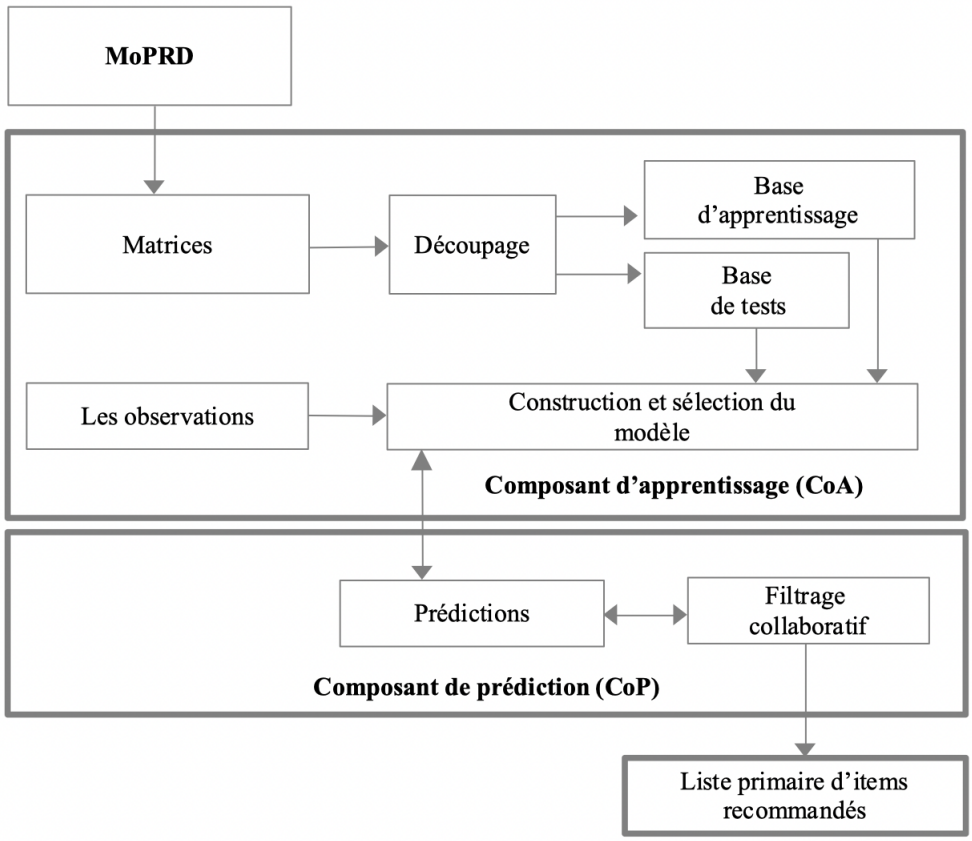

FIGURE 3.11 – Architecture détallée du Module de prédiction

Ce module de prédiction fournit en fin de traitement l'ensemble des items recommandés et obtenus par notre algorithme de prédiction. Pour mieux guider l'expert dans ces choix de recommandation, nous avons exécuté les items recommandés dans un algorithme permettant de faire une classification de ces derniers en fonction de leurs pertinences face aux problématiques des points de vente décelées durant les enquêtes mystères.

3.4 Module de classification des items (MoCI)

Le module responsable de la classification de la liste primaire vient compléter le traitement de notre méthode hybride de recommandation. Le but de cette classification est de formuler une prédiction en créant des classes distinctes dans un ensemble d'items de la liste primaire. Ensuite, sur la base de ces prédictions, l'objectif est de faire une classification pertinente permettant de guider l'expert dans sa prise de décision. En effet, nous

soulignons que le traitement pour la recommandation est partiel et que la décision finale reste toujours humaine. Ainsi, la liste d'items qui est proposée à l'expert ne reste qu'une proposition (Hunt et al., 2019).

Nous proposons dans cette thèse un algorithme permettant de réaliser cette classification des items de la liste primaire sur la base de leur pertinence. Pour ce faire, nous avons utilisé une méthode de classification basée sur l'apprentissage automatique supervisé, notamment l'algorithme des plus proches voisins qui sera détaillé dans le chapitre 4.

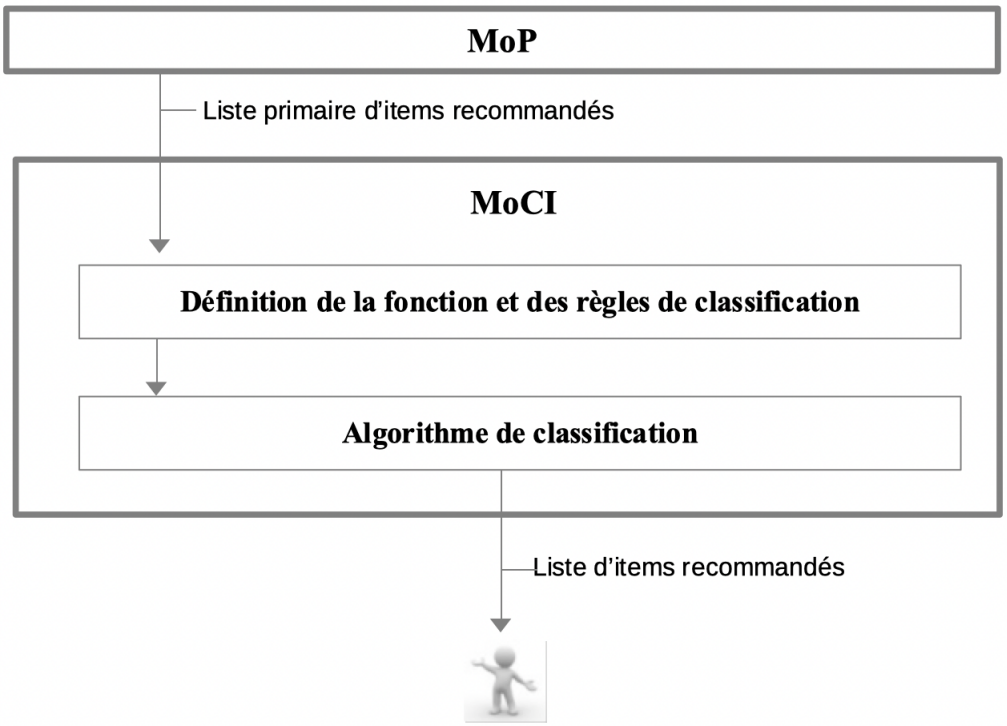

FIGURE 3.12 – Architecture du MoCI

L'apprentissage supervisé consiste en un ensemble de variables d'entrée X et un ensemble de variables de sortie Y. Nous avons eu recours à un algorithme de classification pour apprendre la fonction de mapping de l'entrée à la sortie X. Dans nos travaux, l'entrée correspond à l'ensemble des items présents dans la liste primaire. La sortie Y est la liste secondaire des items après le processus de classification. Ainsi nous avons :

$$Y = F(X) \tag{3.2}$$

L'objectif est de prendre en compte la fonction de mapping quand on a de nouvelles données d'entrées (X), pour que l'on puise prédire les variables de sortie (Y) pour ces données. Le processus est nommé « apprentissage supervisé » , car il est basé sur un algorithme tiré de l'ensemble des données de formation, « training set » peut être considéré comme un enseignant supervisant le processus d'apprentissage. L'algorithme effectue des prédictions itératives sur les données d'entraînement. Il est ensuite corrigé par l'enseignant. L'apprentissage s'arrête lorsque l'algorithme atteint un niveau de performance acceptable.

3.4.1 Définition de la fonction et des règles de classification

La définition de la fonction et des règles de classification est une prémisse pour l'élaboration de l'algorithme de classification. Le but de cette démarche est de définir les besoins pour l'implémentation de l'algorithme permettant de faire le traitement de la classification (Guo et al., 2003). Le résultat obtenu de cette classification est une liste de recommandation destinée à l'expert qui va prendre sa décision définitive pour les propositions d'amélioration qu'il doit fournir aux points de vente.

Il est primordial que les connaissances produites par les points de vente soient intégrées dans le traitement. Les points de vente fournissent une partie des connaissances liées aux domaines dans lesquels le traitement pour la recommandation est appliqué. Si un point de vente détermine le nombre et les identités des classes à obtenir, le choix d'une méthode supervisée serait plus approprié. Cependant, si l'on prend un autre exemple simple qui est celui du nombre de classes souhaité : si un expert désire obtenir un nombre exact de classes sans fournir aucune autre information, dans ce cas la méthode K-means1 sera la méthode la plus adaptée (Likas et al., 2003). Mais si l'expert souhaite un nombre de classes compris entre X et Y, le classifieur peut utiliser l'algorithme ISO data (Myllynen et al., 2021) en lui fixant des paramètres d'initialisation extraits de cette information.

Pour lancer le processus de classification, la fonction d'apprentissage doit être définie et les différentes règles qui vont établir l'algorithme de classement qui sera appliqué à la liste primaire de recommandation. Nous faisons appel à la classification supervisée utilisant la fonction du plus proche voisin (K-NN). La classification supervisée vise à associer chacun

des n observations x_1, ...,x_n à l'un des k classes connues à priori tandis que la classification non supervisée a pour but de regrouper ces données en k groupes homogènes (Guo et al., 2003). Le lecteur pourra trouver de plus amples détails sur ces deux approches dans (Pavlenko & Von Rosen, 2001). À partir de la définition de la fonction et des règles de classification, nous avons proposé un algorithme de classification.

3.4.2 Notre algorithme de classification

Notre algorithme de classification est un deuxième filtrage qui a pour objectif de réduire le temps traitement de l'expert. En effet, en faisant une classification des items qui ont été recommandés par le système, l'expert aura une idée sur les pertinences de ces derniers pour l'amélioration des forces de vente dans le cas d'un réseau de points de vente.

Notre algorithme de classification est construit autour de la recherche de pertinence entre les items proposés dans la liste primaire afin de générer une liste secondaire d'items. Cette liste secondaire est un classement par ordre de pertinence sur la base des besoins d'amélioration en marketing des points de vente qui ont sollicité l'enquête. L'algorithme de classification se base ainsi sur les concepts qui ont été évalués. Ces concepts représentent les besoins qui ont été capturés et jugés pertinents à évaluer durant la phase de collecte qui a eu lieu en début de traitement pour la recommandation. Notre algorithme de classification est construit sur la base d'un ensemble comportant :

- les concepts issus des observations à évaluer concernant le point de vente qui a sollicité le traitement ;
- les différentes règles issues du domaine.

Notre algorithme est orienté vers un regroupement des items qui sont basés sur les règles et les concepts qui ont été dans l'ontologie du domaine. Ainsi, notre algorithme est formulé comme suit :

Algorithm 2 Notre algorithme de classification

Données en entrée :

• La liste primaire d'items C.

• L'ensemble des données sélectionnées D.

• L'ensemble des règles issues du domaine R

• La fonction permettant de définir la distance d.

Soit X appartenant à la liste primaire d'items C. Pour obtenir la

prédiction de X en une sortie y on fait :

• calculer l'ensemble des distances de l'observation X sous la base de R, avec les autres observations de l'ensemble des données D et de la liste primaire C ;

• sauvegarder les K observations de l'ensemble des données D les proches de X en utilisant la fonction de calcul de distance d ;

• prendre les valeurs de y des D observations sauvegardées : on effectue une classification , calculer le mode de y retenus ;

• retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par K-NN pour X.

3.5 Contributions

Nos contributions permettent d'automatiser le processus de recommandation, d'apporter des améliorations sur le traitement de données et de proposer des recommandations. La contribution majeure apportée au logiciel de Retaily est notre méthode de recommandation combinant les technologies du web sémantique et celles de l'apprentissage automatique. Le SRSE que nous proposons apporte plusieurs contributions par rapport à la littérature existante en particulier l'homogénéisation des données et l'amélioration du démarrage à froid. En effet, nous proposons dans cette thèse une méthode permettant l'amélioration du démarrage à froid dans le contexte d'un nouveau point de vente qui rejoint le système. La Figure 3.13 illustre l'architecture globale détaillée de notre SRSE.

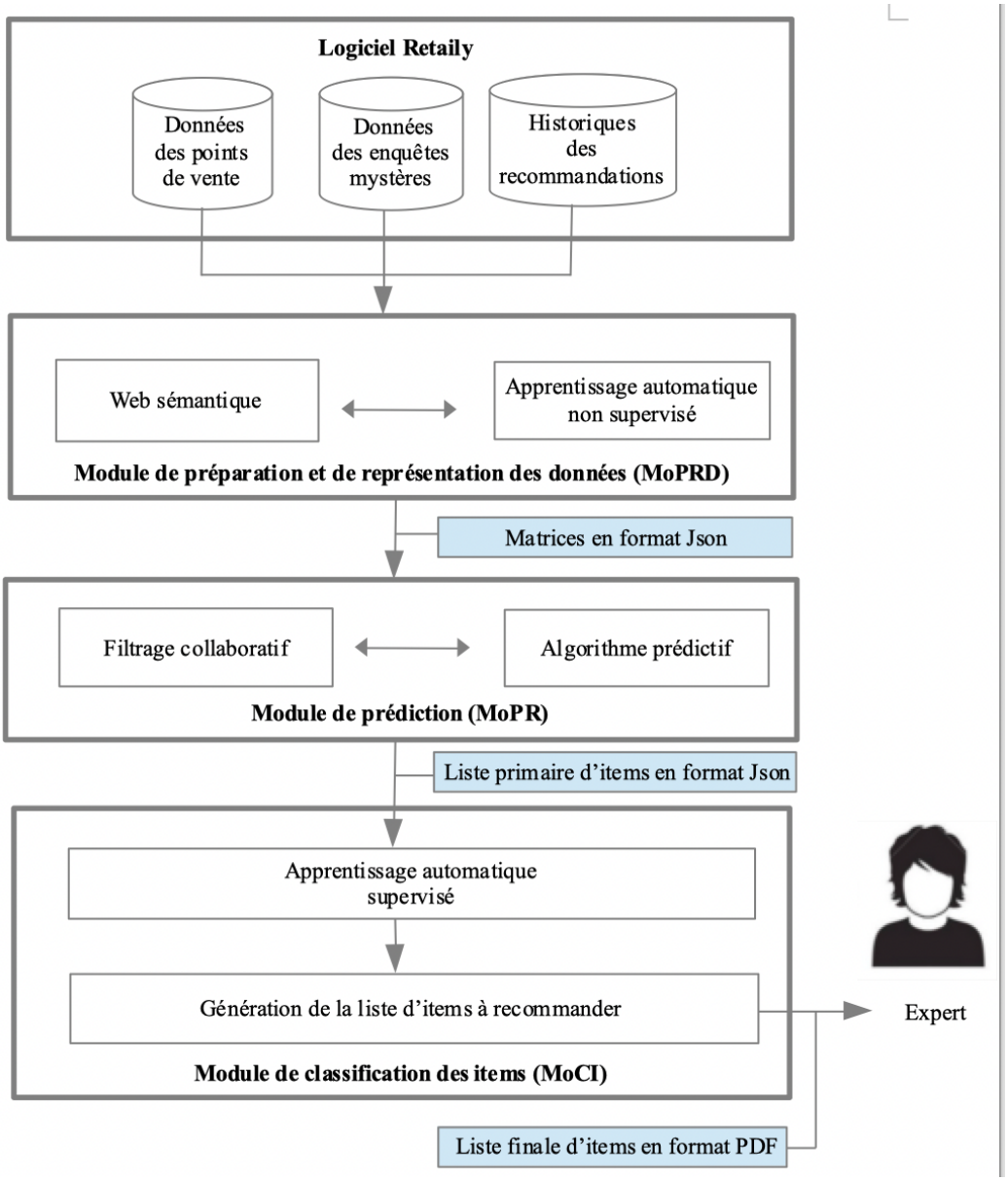

FIGURE 3.13 – Architecture globale détaillé

*

* *

Dans ce chapitre, nous avons abordé, la modélisation de la méthode hybride de notre système de recommandation sémantique enrichi (SRSE). Dans un premier temps, une présentation de l'architecture globale du système a été effectuée. Ensuite, les différents modules de notre architecture sont présentés dans les détails et les contributions résultantes de notre proposition ont été exposées. Enfin, la définition des différents axes d'évaluations de ces modules et de notre système dans sa globalité sont résumés dans une grille de lecture. Ces axes d'évaluations vont être développés et utilisés dans le chapitre suivant pour la validation de nos contributions après expérimentations.

Chapitre 4

Implémentation et évaluation

Sommaire

4.1 Implémentation

« *Le meilleur moyen de prédire le futur est de l'inventer.* »

Alan kay

L'implémentation de notre SRSE s'appuie sur la modélisation qui a été réalisée dans le chapitre 3. Pour rappel, notre système de recommandation sémantique enrichi (SRSE) est décomposé en trois modules. Ces trois modules vont être implémentées de manière indépendante. Cette technique d'implémentation a été choisie dans le but de pourvoir réutiliser les différents modules dans d'autres projets. Ainsi, pour chaque module nous aurons un ensemble de données d'entrée et un ensemble de données de sortie. Les différents programmes informatiques de ces modules ont été implémentés en Python, ce choix est expliqué dans la section suivante.

4.1.1 Les technologies impliquées dans notre proposition

Dans notre implémentation, nous avons utilisé l'API de TensorFlow pour une partie du traitement, notamment pour faire appel aux technologies de l'apprentissage automatique. Des ontologies concernant nos domaines d'applications ont été exploitées pour la partie utilisant les technologies du web sémantique.

TensorFlow

TensorFlow est une bibliothèque de l'apprentissage automatique gratuite, créée par Google. Elle permet de développer et d'exécuter des applications d'apprentissage automatique et d'apprentissage profond « Deep Learning ».
L'apprentissage automatique est très utile dans de nombreux cas d'usage comme explicité dans le chapitre 1, page 24, mais malheureusement complexe à mettre en place (Géron, 2019). Son usage est utilisé pour :

- collecter des données ;
- entraîner des modèles ;
- réaliser le déploiement de réseaux de neurones, qui requiert à l'origine d'importantes compétences techniques.

Il existe plusieurs bibliothèques d'apprentissage automatique, le Tableau 4.1 présente quelques-unes parmi elles et illustre une comparaison entre ces dernières (Gevorkyan et al., 2019).

	Plateforme	Coût	Écrit en langue	Algorithme ou fonctionnalités
Ensemble	Linux, Mac OS, Windows	Gratuit	Java	- Préparation des données - Classification - Régression - Clustering - Visualisation - L'association règle l'exploitation minière
PyTorch	Linux, Mac OS, Windows	Gratuit	Python, C ++, MIRACLES	- Module Autograd - Module optimal - Module nn
Scikit Learn	Linux, Mac OS, Windows	Gratuit	Python, Cython, C, C ++	- Classification - Régression - Clustering - Prétraitement - Sélection de modèle - Réduction de dimensionnalité.
Google Tensor-Flow	Linux, Mac OS, Windows	Gratuit	Python, C ++, MIRACLES	Fournit une bibliothèque pour la programmation de flux de données.

TABLE 4.1 – Comparaison de quelques bibliothèques d'apprentissage automatique

Nous avons choisi d'utiliser Tensorflow pour trois raisons. La première est que nous la considérons plus facile pour un débutant ou pour un expert de créer des modèles d'appren-

tissage automatique. La deuxième raison est la facilité qu'elle offre pour le déploiement des modèles d'apprentissage sur le cloud, sur des sites web ou encore sur des appareils. La dernière raison est son écosystème qui est complet pour nous aider à utiliser avec efficacité les outils d'apprentissage automatique.

Les ontologies utilisées dans cette thèse

Les deux ontologies que nous avons utilisées dans cette thèse sont issue des domaines de gestion de compétences des personnels d'une organisation et celle de la structuration d'une organisation. La notion d'organisation correspond dans notre cadre d'étude, à un point de vente, ou encore une entreprise. Ces deux ontologies sont :

1. une première ontologie sur la gestion des compétences de Fazel-Zarandi et Fox (2012). Cette ontologie permet de représenter, d'inférer et de valider des compétences des employés d'une entreprise au fil du temps. la modélisation des ressources humaines dans un environnement dynamique. Cette ontologie permet de spécifier les compétences à des niveaux particuliers de compétence comme ce qui permet de décrire la performance des activités, et les énoncés de compétences en tant que propriétés auxquelles sont associés des degrés de croyance et qui peuvent changer avec le temps. La Figure 4.1 est une illustration de la taxonomie de l'ontologie de Fazel-Zarandi et Fox (2012) . Nous avons choisi d'utiliser dans nos expérimentations cette ontologie, car étant en adéquation avec notre cadre d'étude, notamment à la conceptualisation des compétences des employés d'un réseau de points de vente ;

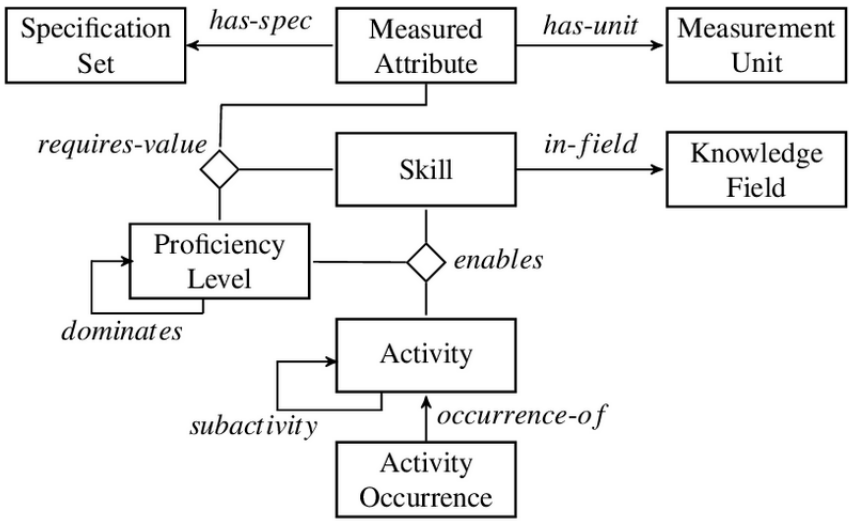

FIGURE 4.1 – Taxonomie de l'ontologie de la gestion des compétences (Fazel-Zarandi & Fox, 2012)

2. la deuxième ontologie associée à notre représentation sémantique est en charge de la représentation conceptuelle des différents points de vente. Cette ontologie a été développée par Fox et al. (1996) et permet de lier la structure et le comportement d'une organisation (d'un point de vente dans notre cas d'étude). Selon Fox et al. (1996) l'objectif du projet de modélisation des organisations est de créer le modèle d'entreprise de nouvelle génération. La Figure 4.2 illustre une partie de la taxonomie de l'ontologie de Fox et al. (1996).

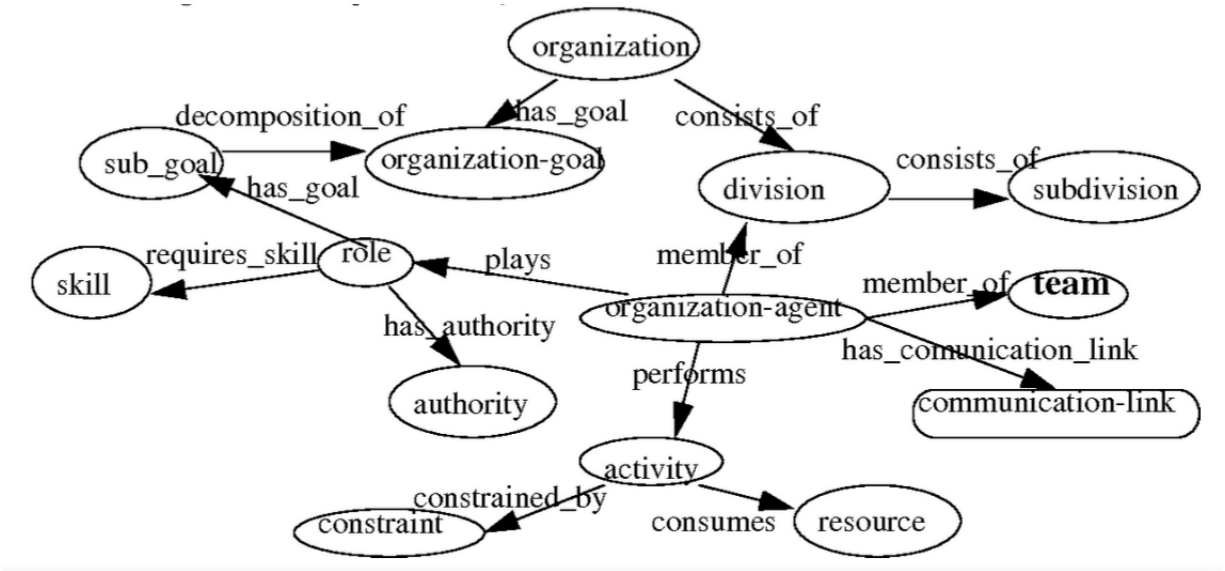

FIGURE 4.2 – Taxonomie de l'ontologie du comportement d'une organisation (Fox et al., 1996)

Après avoir présenté les différentes technologies qui nous ont permis de mettre en place un prototype de notre SRSE, nous allons expliciter son fonctionnement.

4.1.2 Fonctionnement de notre SRSE

Comme le montre, le diagramme d'activité, illustré dans la Figure 4.3, le fonctionnement de notre SRSE commence par la sélection des données à partir de Retaily. Cette sélection de données est effectuée sous deux formes :

1. active, pour les données sélectionnées de manière explicite. Ce sont les données obtenues explicitement c'est-à-dire celles sélectionnées à partir des résultats des enquêtes mystères ;
2. passive, c'est-à-dire dans le cas où la collecte des données a été réalisée de manière implicite. Nous sommes dans le contexte, où les données ont été obtenues par un ensemble d'observations (voir chapitre 3, page 104) ou par des analyses faites sur les points de vente, par exemple le taux de fréquentation en fonction des tranches horaires.

Après la sélection des données, on passe à l'extraction et à la représentation des connaissances. L'extraction des connaissances, est un processus qui se déroule suivant une suite

d'opérations tel que vu dans le chapitre 2, aux pages 71-75. L'extraction de connaissances à partir de données va nous permettre de passer à l'étape de la représentation des connaissances qui a pour objectif de mettre du sens entre ces différentes connaissances dans un domaine précis. Pour ce faire les ontologies présentées précédemment ont été utilisées dans notre cadre d'étude pour réaliser cette tâche et faire la formation de communautés de points de vente vu dans le chapitre 3, à la page 105.
Ensuite, on passe aux étapes de prédiction et de classification des items pour finir à la restitution des résultats à l'expert.

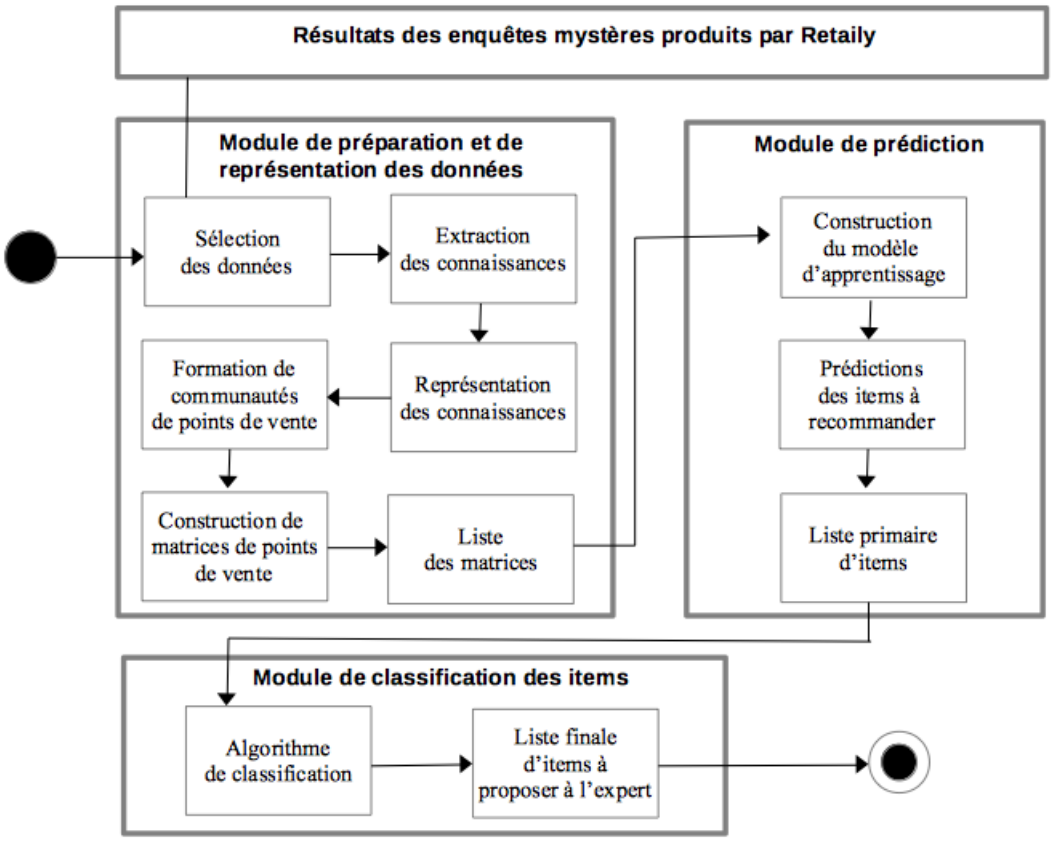

FIGURE 4.3 – Diagramme d'activité

Sélection des données pour la recommandation

Nous avons travaillé sur les données obtenues des résultats d'enquêtes mystères pilotées par le logiciel Retaily. Ces données d'enquêtes mystères sont recueillies via des questionnaires qui ont été prédéfinis préalablement par les commanditaires sur la base

d'un ensemble de critères d'évaluation. Ces critères d'évaluations représentent aussi les différents items que nous allons retrouver dans nos matrices de recommandations et dans la liste des items qui seront proposés à l'expert pour sa décision finale. En plus d'être les items dans notre processus de recommandation, ces critères d'évaluations sont traduits dans la représentation ontologique en des concepts. En plus des données d'enquêtes mystères, les données liées directement aux points de vente qui ont été évaluées sont aussi sélectionnées. Cela donne lieu à deux manières de collecter des données :

1. Les données obtenues à partir des enquêtes mystères et les données qui ont été fournies par les points de vente de manière explicite. Elles reposent sur le fait que les clients mystères indiquent explicitement leurs appréciations au système avec des actions comme commenter, taguer/étiqueter ou encore noter. On utilise souvent une échelle de ratings allant de 1 étoile à 5 étoiles qui est par la suite transformées en valeurs numériques afin de pouvoir être utilisée par les algorithmes de recommandation. L'avantage de cette manière de collecter des données est la capacité à reconstruire l'historique d'un point de vente et la capacité à éviter d'agréger une information qui ne correspond pas à cet unique point de vente (plusieurs points de vente dans un même réseau ou dans un même secteur). Son inconvénient est que les données recueillies peuvent contenir un biais dit de déclaration ;
2. La sélection des données obtenues à partir d'observations (la notion d'observation a été abordée dans le chapitre 3, à la page 104) et sur des analyses de comportements des points de vente effectuées de façon implicite dans l'application qui embarque le système de recommandation, cela se fait en arrière-plan (globalement sans rien demander aux responsables des points de vente). Par exemple, inclure dans la sélection les données sur les statistiques de vente à partir du site du point de vente ou encore obtenir les données sur le taux de fréquentation du point de vente en fonction de tranches horaires via une plateforme tiers.

   Nous avons profité ici de l'avantage de cette manière de collecter des données implicites, car aucune information n'est demandée aux responsables des points de vente, toutes les données sont collectées automatiquement. Par contre, nous avons hérité de certains inconvénients sur les données récupérées, car ces dernières sont plus difficilement attribuables à un point de vente et peuvent donc contenir des

biais d'attribution (utilisation commune d'un même compte par plusieurs points de vente). Deux points de vente appartenant à un même réseau peuvent ne pas avoir les mêmes taux de fréquentation, ou ne pas avoir les mêmes profils des clients en termes d'âge.

Après la phase de sélection de données, des connaissances sont extraites de ces dernières pour former notre base de connaissances. Comme nous l'avons présenté dans le chapitre 2, dans une base de connaissance qui regroupe l'ensemble des instances (A-Box) des concepts définis dans une ontologie du domaine donné. Cela correspond à l'ensemble des faits ou des objets associés à des concepts durant la phase de peuplement. Ils sont donc catégorisés en fonction de la structure de l'ontologie. Pour permettre à notre SRSE de recommander des items appartenant à un domaine précis, noua avons associé nos items aux éléments de l'ensemble des faits. Pour ce faire, nous avons procédé de deux manières :

1. une mise en correspondance directe et l'assimilation des items entre eux-mêmes à des instances ;
2. une mise en correspondance indirecte et le stockage des items dans une base de données en conservant les références liées aux instances.

Le fait que nous considérons les items au sein de notre SRSE comme des instances d'une ontologie du domaine possède des avantages significatifs. En effet, en procédant ainsi, une classification des items est systématiquement établie grâce à la structuration des concepts auxquels ils s'associent. A l'opposé d'un système de recommandation ne possédant d'aucune information sémantique sur les items, un tel système a l'avantage de disposer de connaissance sur les items. Cependant, certains de nos raisonnements logiques peuvent être appliqués grâce aux relations sémantiques du modèle. Ces raisonnements sont des inférences sur les bases de connaissances. Il est également possible d'y appliquer certaines mesures sémantiques et ainsi de considérer que deux items sont proches sur la base des résultats de nos mesures de similarité sémantique. Nos mesures peuvent être d'une grande finesse si l'ontologie du domaine sur la-quelle repose le système est définie de façon précise et détaillée. Ainsi, selon les compétences recherchées exprimées par un point de vente pour un item particulier (ou un ensemble d'items), des items considérés comme étant proches à l'aide de ces mesures sont pour nous des recommandations à prendre en compte. Nous

nous sommes basés sur cette démarche pour améliorer le démarrage à froid de nouveaux items. En effet, tel que vu dans le chapitre 1, quand un item n'est pas encore noté et qu'on ne connaît donc pas le vecteur de notes qui lui est associé, il est tout de même possible de le recommander. Nous abordons plus bas nos mesures de similarités sémantiques des items au sein de notre SR. Dans notre cadre d'étude, pour qu'un point de vente reçoit en recommandation des items obtenu sur la base de similarité sémantique, il est nécessaire que ce point de vente soit modélisé par les éléments de l'ontologie.

Modélisation du profil des points de vente par des éléments d'une ontologie

Dans notre base de connaissances, les instances sont caractérisées par les concepts de l'ontologie pour lesquels elles sont rattachées et par les attributs liés à ces concepts. Dans notre cas d'étude, la vente d'automobile, une instance du concept secteur d'activité possédant des attributs tels que : modèle, gamme, Prix, etc.
Durant le peuplement de l'ontologie, nous avons associé les instances aux concepts pertinents qui sont les plus bas dans l'arborescence, i.e. les concepts les plus spécifiques. Les relations qui existent entre une instance et des concepts plus génériques, ancêtres de ceux qui lui sont associés ne sont pas explicités, car ils peuvent être inférés par le raisonnement. Nos différents profils de points de vente visent à caractériser les préférences de ces derniers relativement aux compétences pour lesquelles ils se sont déjà exprimés, soit par des préférences explicitées à partir de notes. Ainsi, les profils des points de vente sont sous la forme d'un vecteur dont chaque indice se réfère à un concept et dont la valeur représente le taux d'intérêt du point de vente pour ce concept. En d'autres termes, nous pouvons considérer que les profils de nos points de vente sont représentés par les éléments présents dans la T-Box associés à différents poids selon les intérêts exprimés. Dans notre méthode, le profil d'un point de vente est représenté ici comme un ensemble de nœuds dont chacun est représenté sous forme de paire (Co_i, $IS(Co_i)$) (avec Co, un concept défini dans l'ontologie et $IS(Co_i)$, le taux d'intérêt d'un point de vente pour Co_i). Nous avons utilisé par la suite ces profils pour estimer la similarité entre les points de ventes dans une approche hybride. Nous utilisons ce formalisme en nous basant sur les travaux de Middleton et al. (2004) qui adoptent des triplets,(utilisateur, thème, taux d'intérêt) pour modéliser l'intérêt des utilisateurs pour des articles scientifiques de différents domaines.

Inférence du taux d'intérêt et mise à jour du profil d'un point de vente

L'initialisation des différentes valeurs du vecteur correspondant au degré d'intérêt de nos points de vente est importante puisqu'elle permet de proposer des recommandations dès la première connexion au SRSE. Pour ce faire, en amont de la recommandation, nous avons récupéré les données fournies par les points de vente lors de leurs inscriptions. Au lieu de questionner les points de vente sur leurs compétences recherchées par rapport à chaque concept de l'ontologie pour engendrer un profil initial complet, nous nous sommes basé sur les travaux de Moreno et al. (2013). Ils proposent dans leur Système de recommandation des activités touristiques, de ne questionner l'utilisateur que sur quelques concepts généraux, mais suffisamment significatifs pour représenter les principaux centres d'intérêt des touristes (e.g. Plage, Shopping, Culture, Gastronomie, etc.).
L'atout de demander directement aux points de vente d'exprimer leurs compétences recherchées pour construire leurs profils est que cela permet d'obtenir des recommandations précises. Cependant, cela demande un effort aux points de vente et cette activité en plus d'être chronophage peut également être perçue comme trop intrusive.
Paradoxalement, dans la plupart des cas, les points des ventes cherchent à recevoir des recommandations précises et pertinentes, mais sans trop se dévoiler au système, sans être interrogés de façon précise.
Dans notre cas d'étude la formation de communautés de points de vente (voir chapitre 3, page 105) et les résultats d'enquêtes mystères vont venir compléter et préciser les recommandations. Une autre solution consiste à attribuer initialement le même poids à chaque concept du profil du point de vente. Mais ce n'est qu'au fur et à mesure de l'utilisation du SR que le profil de chaque point de vente sera affiné et que les recommandations gagneront en pertinence.

Calcul de similarité et formation de communautés de points de vente

Pour réduire le temps de traitement de notre SRSE, nous avons intégré dans notre approche la création de communautés de points de vente. Cette notion de groupement de points de vente dans un SR a été explicitée dans le chapitre 3, page 105. L'objectif premier de cette démarche est d'optimiser la performance de l'algorithme de recommandation.

Elle va aussi nous permettre d'améliorer le démarrage à froid. Comme souligné dans le chapitre 3 à la page 95, une contribution sur l'amélioration du démarrage à froid pour un nouvel utilisateur est proposée dans cette thèse. La formation de communautés de points de vente est au centre de cette proposition d'amélioration.
Notre technique de formation de communautés de points de vente est basée sur le calcul de la similarité sémantique entre les différents points de vente du SRSE. Pour ce faire nous faisons appel à deux types de métriques, la première métrique permet de faire le calcul de la similarité sémantique entre les points de vente à partir de leurs profils et la seconde à partir des concepts obtenu de l'extraction des connaissances et de la représentation de ces dernières. Ensuite, nous allons comparer les résultats obtenues de ces méthodes de calcul de similarité.
Comme présenté dans le chapitre 1, la recherche de la similarité entre deux utilisateurs permet de rechercher ceux qui sont très proches en termes de préférences ou d'historique d'achats ou encore dans notre cas en termes de profil utilisateur. En effet, nous nous intéressons beaucoup à la recherche de la similarité entre les différents profils utilisateurs. Cette recherche de similarité basée sur les données du profil utilisateur est une démarche permettant d'améliorer le démarrage à froid pour un nouvel utilisateur. Après la formation des communautés d'utilisateurs, nous passons à la génération de l'ensemble des matrices nécessaires pour le traitement de la recommandation.

Matrices de recommandation

Nos matrices de recommandation représentent la structure d'entrée de l'ensemble des données vers notre algorithme de prédictions que nous allons voir par la suite. Les données représentées dans l'ensemble des matrices sont les données qui sont propres aux points de vente et les résultats des enquêtes mystères.
Nous utilisons dans notre SRSE le filtrage collaboratif (FColl), ainsi notre système ne prend en compte que les appréciations associées à ses items. Typiquement, comme chaque point de vente ne possède une note que sur une partie des items, ces notes engendrent une matrice incomplète, appelée la matrice (pointsDeVente-items), dans laquelle chaque ligne représente un point de vente.
L'algorithme de prédiction permet de compléter cette matrice en inférant les données

manquantes. Afin de prédire la note qu'un point de vente pourrait obtenir sur un item, le système détermine d'abord un ensemble des points de vente similaires, ses voisins. La similarité sémantique entre points de vente est calculée en mesurant la similarité de leurs vecteurs de notation.

Prédictions et traitement pour la recommandation

La phase de prédiction est incontournable pour un moteur de recommandation. Cette phase est au cœur du traitement d'un système de recommandation. C'est une étape qui va nous permettre de prédire l'utilité de chaque item pour un point de vente sollicitant la recommandation. Dans cette démarche de prédiction nous avons choisi l'utilisation des technologies d'apprentissage automatique, et précisément l'algorithme de K-NN, présenté dans le chapitre 3. C'est un algorithme qui n'a pas besoin de modèle pour faire des prédictions. Mais pour le rendre plus efficace nous avons inclus la notion de modèles d'apprentissage. Notre algorithme K-NN est divisé en plusieurs étapes :

1. la première étape consiste à importer l'ensemble des « data-sets ». Pour faciliter le traitement de l'ensemble de nos « data-sets » est importé dans un fichier CSV. Ces « data-sets » sont toutes les données que nous avons dans les matrices de recommandation que nous avons présentées précédemment. Pour réaliser l'import nous avons utilisé l'api de Tensorflow qui utilise les bibliothèques Python requises telles que Pandas, Numpy, Seaborn ou encore Matplotlib. Ensuite les fichiers CSV sont importés à l'aide de la fonction *read_csv* prédéfinie dans Pandas (voir algorithme d'importation des data-sets dans l'annexe A) ;

2. en deuxième étape, Nous avons d'abord utilisé la fonction head(), describe() pour afficher les valeurs et la structure de l'ensemble de données, puis procéder au nettoyage des données. Cette deuxième étape est une étape d'exploration et de nettoyage des « data-sets » mais aussi de découpage. Ce découpage consiste à faire la séparation en trois groupes de « data-sets ». Un premier groupe de données de test, un deuxième groupe de données d'entrainement et un dernier groupe de données d'observations. Les données de test et d'entrainement vont nous permettre de générer nos modèles d'apprentissage. Les données d'observation nous

permettent de faire une mise à jour de nos modèles et de les rendre plus pertinents (Voir dans l'annexe A, algorithme de lecture et de concaténation des données d'entraînement) ;

3. en troisième étape, notre algorithme va procéder à l'entrainement avec nos différents modèles d'apprentissage à l'aide des données de test et d'entrainement. Ensuite à partir des observations il va sortir les premières prédiction sur la base de ces modèles. Ces derniers sont construits à partir des métriques que propose l'algorithme de KNN. Les observations représentent dans notre cas d'étude, l'ensemble des derniers résultats d'enquête mystère. Cette étape est une étape de mise à jour de nos modèles et aussi d'entrainement pour nos modèles. Pour la démarche d'entrainement nous avons utilisé l'API de Tensorflow ;
4. après l'obtention de la première liste de prédictions à partir des observations, un premier filtrage est effectué pour éviter les redondances de prédictions. Pour ce faire nous avons développé un module de filtrage automatique. Ce dernier est un algorithme très simple basé sur la recherche de similarité sémantique entre les items afin de supprimer ceux qui sont redondants. Ce module de filtrage conclut notre processus de prédiction en proposant en sortie une liste primaire de recommandations. Cette liste primaire contient les données d'entrée du module suivant qui est en charge de la classification des items.

Classification des items recommandés

Notre classification d'items à recommander est une contribution qui va nous permettre de guider l'expert en charge de la prise de décisions finales. Elle représente la dernière étape de traitement de notre processus de recommandation avant la prise de décision finale par l'humain. En effet, notre contribution d'automatisation d'analyse et de recommandation de plans d'action est partielle.

Les technologies d'apprentissage automatique sont utilisées ici pour réaliser la classification des items à recommander. En effet, ces technologies proposent une démarche de classification basée sur l'apprentissage supervisé comme présenté dans le chapitre précédent. Ce module va recevoir en entrée la liste primaire des items issus du module de prédiction. Dans notre cas d'études, il s'agit de classer une liste primaire de compétences destinée

à un ou plusieurs points de vente dans le but d'améliorer leurs forces de vente. Pour ce faire nous avons fait appel à l'API de Tensorflow et notre démarche de classification est effectuée comme suit :

1. collecte et exploration des data-sets, ici les « data-sets » issus de la liste primaire d'items sont considérés comme des données d'observations, ensuite l'historique des recommandations déjà reçues et validées comme pertinentes par le point de vente dans le passé constituent les données d'entrainement. Enfin, les données de test sont l'ensemble des données de l'historique du point de vente (Voir l'algorithme de classification des items présent dans l'annexe A). Après cette phase de collecte et d'exploration des data-sets, on passe à l'étape d'entraînement ;
2. l'entrainement et les fonctions d'apprentissage : avant de passer à l'entrainement on peut consulter les statistiques de notre quantité de données à lancer pour l'entrainement et ensuite procéder à la sélection des colonnes (voir l'algorithme d'entraînement du modèle et l'algorithme de prédiction dans l'annexe A) ;
3. la proposition de la liste secondaire et finale des items à recommander à l'expert est obtenue à partir du calcul des scores de l'ensemble des prédictions (voir l'algorithme de classification finale dans l'annexe A).

Présentation de la liste des items recommandés

Pour mieux guider l'expert qui est en charge de la prise de décision finale, la manière de présenter la liste des items recommandés est très importante. En effet pour une lecture facile des items recommandés par notre algorithme, nous avons modélisé une fiche de lecture pour l'expert. La grille de lecture illustrée dans le Tableau 4.2 est est un exemple et est caractérisée par un rang, taux de pertinence, nombre de récurrences à la recommandation et degrés d'importance pour le réseau de points de vente sollicitant la recommandation.

Rang	Item	Nombre de récurrences	Taux de pertinence	Degrés d'importance
1	Courtoisie	7	80	80
2	Rapidité	27	67	60

TABLE 4.2 – Exemple d'une grille de lecture d'une liste d'items recommandés proposée à l'expert

Cette grille de lecture des résultats de la recommandation concerne, un point de vente. Ainsi chaque point de vente obtient sa propre grille de lecture des recommandations produites par SRSE. Cette grille de lecture correspond aussi aux résultats de nos hypothèses de recherche exécutées à travers un prototype informatique qui est notre SRSE. Pour valider nos différentes hypothèses, nous abordons dans la section suivante, l'évaluation de notre SRSE.

4.2 Évaluations du SRSE

Un ensemble d'évaluations a été réalisé pour valider nos différentes hypothèses présentées en détails dans le chapitre 3. Les données d'évaluation qui ont été utilisées dans cette thèse sont les résultats des enquêtes mystères.
Dans un premier temps, pour mieux comprendre le traitement des données d'enquêtes mystères, nous allons présenter les modèles de données utilisés par le logiciel Retaily, ensuite les résultats obtenus après expérimentation de notre SRSE sont présentés et enfin une discussion entre ces résultats et les travaux existants sera réalisée.

4.2.1 Modèles de données du logiciel Retaily

Pour illustrer la structuration des modèles de données du logiciel Retaily, nous allons présenter les diagrammes relationnels, diagramme de classe et de séquence.

La Figure 4.4 montre la structure de la base de données, notamment du module en charge de la gestion des enquêtes mystères du logiciel de Retaily. Ce module de gestion des enquêtes mystères est composé d'une relation de 11 entités. Ces entités permettent de stocker les données de la création de la campagne à la restitution des rapports générés par le logiciel après les enquêtes. Le logiciel Retaily propose deux types d'enquête : une enquête de type visite mystère et une autre de type satisfaction qui est réalisée à la suite d'une visite mystère. L'enquête de satisfaction permet d'évaluer l'impact des modifications d'amélioration qui ont été apportées après une visite mystère au sein d'un point de vente.

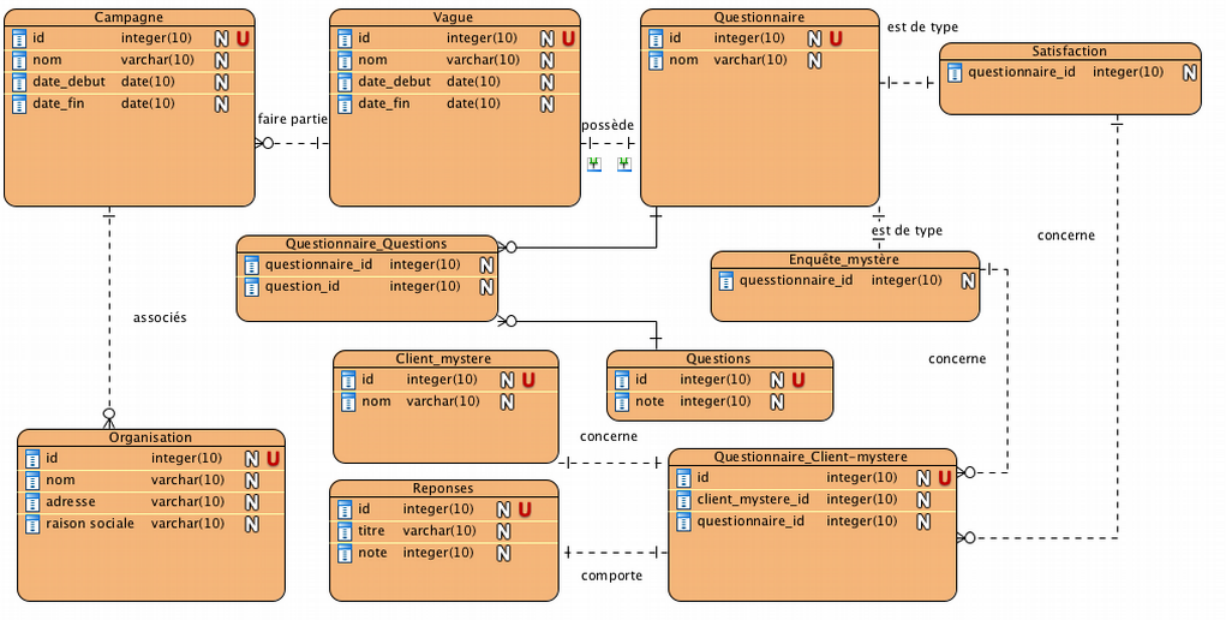

FIGURE 4.4 – Diagramme relationnel

Nous présentons respectivement à la Figure 4.5 et 4.6, le diagramme de classe et le diagramme de séquence. Ce diagramme de classe vient soutenir la compréhension de la structure et le traitement des données d'enquêtes mystères avec le logiciel Retaily.

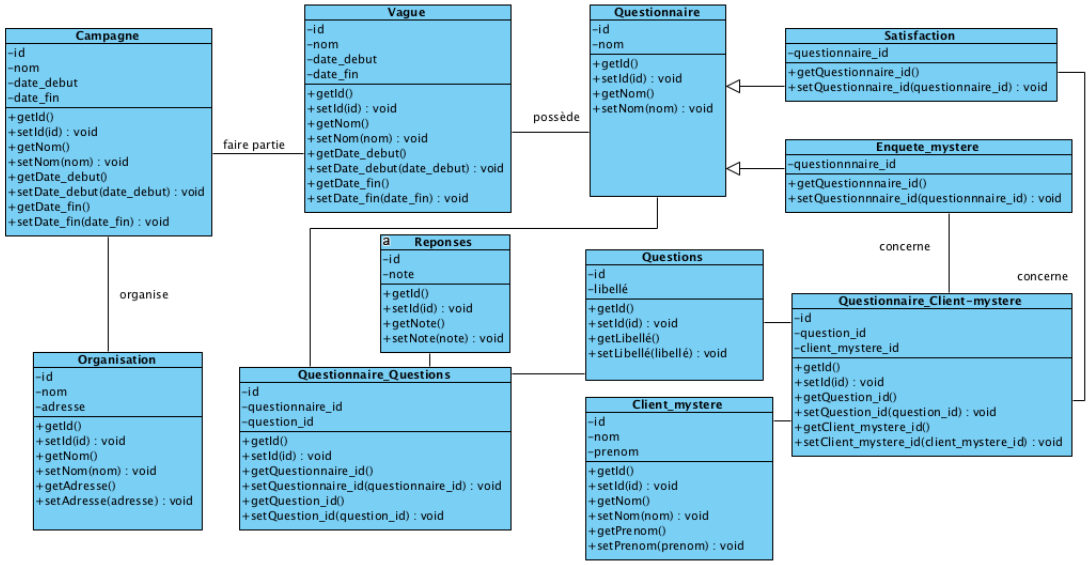

FIGURE 4.5 – Diagramme de classe

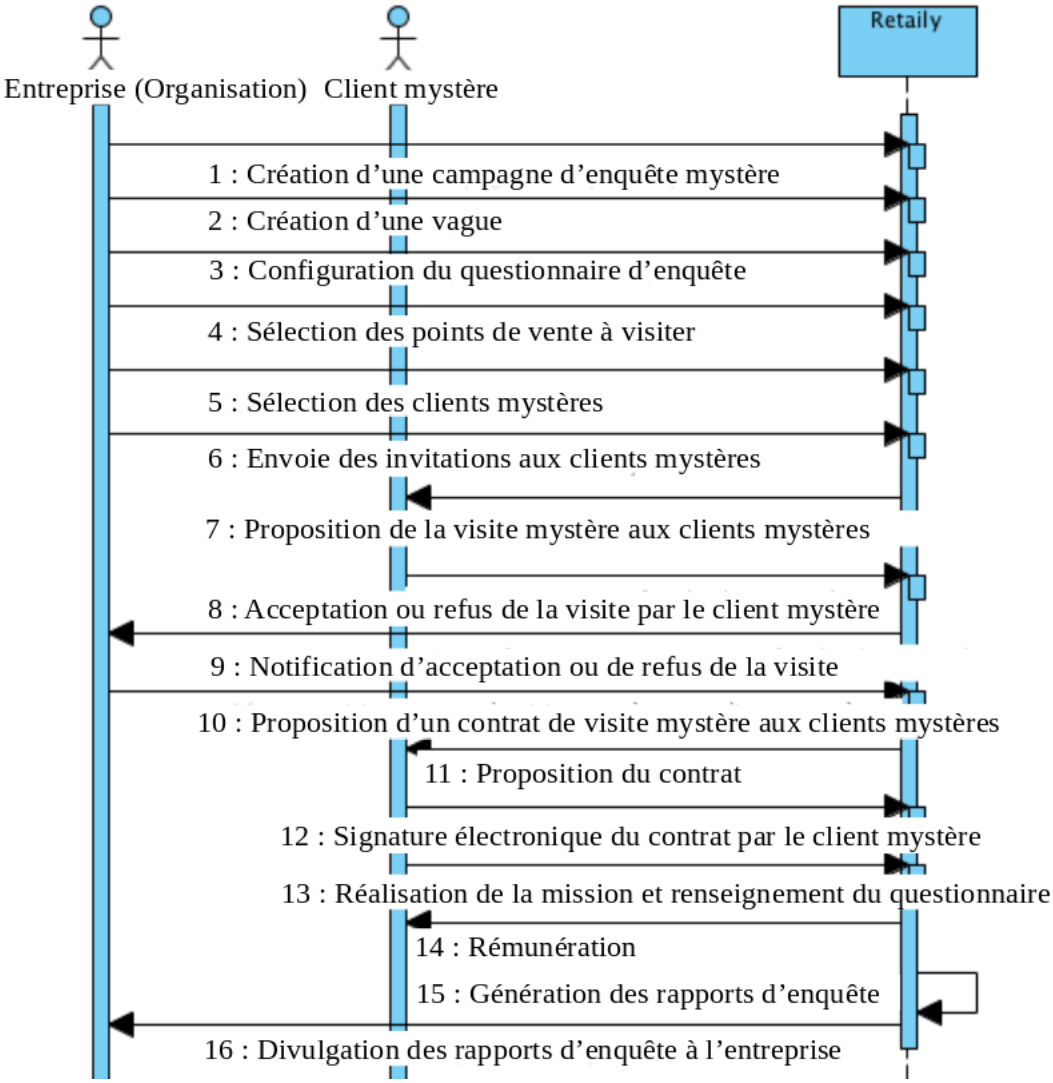

FIGURE 4.6 – Diagramme de séquence

En rajoutant notre SRSE à la démarche de traitement du logiciel des enquêtes mystères, nous obtenons le diagramme de séquence illustré dans la Figure 4.7.

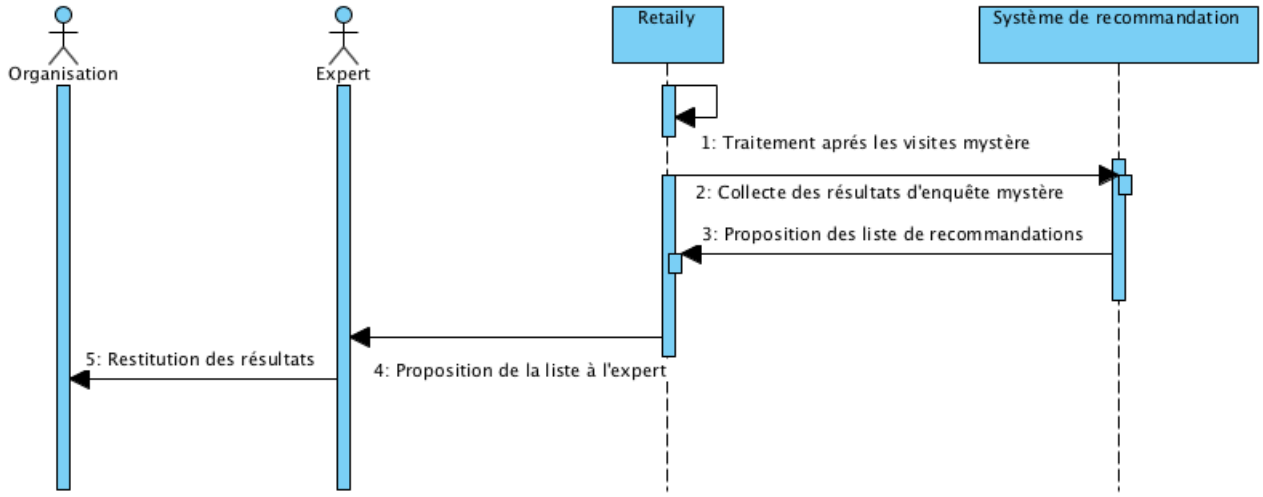

FIGURE 4.7 – Diagramme de séquence avec un SR

4.2.2 Méthodes d'évaluation

Le Tableau 4.3, montre les différentes méthodes d'évaluation qui ont été appliquées à notre SRSE. Dans ce tableau nous montrons les différentes hypothèses à évaluer en indiquant les modules où elles interviennent dans notre architecture globale de contribution. Ensuite, nous indiquons les techniques qui sont utilisées pour les évaluer et les valider. En début de thèse, les évaluations de ces hypothèses étaient prévues en trois phases :

1. Le calcul de la pertinence des items recommandés avec la mesure du score F [1] et du « Recall [2] » de Herlocker et al.(2004).

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.1}$$

Pour obtenir les résultats liés à l'homogénéisation des données, nous avons utiliser les métriques permettant de chercher le nombre de relations entre les différentes données collectées. Ces métriques sont proposées par « Page Rank » de Google,

1. Le F-Score ou F-mesure est une mesure de la précision d'un test statistique. Il tient compte à la fois de la précision et des mesures de « Recall » du test pour calculer le score. On pourrait l'interpréter comme une moyenne pondérée de la précision et de la mémoire, où le meilleur score F1 a sa valeur à 1 et le pire score à la valeur 0. Dans le domaine des recommandations, elle est considérée comme une valeur unique obtenue en combinant à la fois les mesures de précision et de rappel et indique une utilité globale de la liste de recommandations.
2. Le « Recall » mesure la proportion de tous les résultats pertinents inclus dans les meilleurs résultats.

qui permet de chercher le nombre de relations que possède une page web. C'est une technique efficace pour trouver le nombre de relations qu'un concept possède au sein d'une ontologie (Jones & Alani, 2006 ; Sicilia et al., 2012 ; Syed et al., 2010). Soit Pre(P_i) la situation précédente par rapport aux relations du concept et L(P_i), le nombre de relations sortantes. Pr(P_i), le nombre de relations sortantes est exprimé par la formule suivante :

$$Pr(P) = \sum \frac{Pre(P_i)}{L(P_i)} \tag{4.2}$$

2. L'évaluation de la liste de recommandations par un expert en marketing ;
3. Une évaluation auprès des magasins sur les recommandations proposés par l'expert à partir de la liste de recommandation de notre système.

Avec la situation sanitaire liée à la pandémie du COVID 19, les évaluations auprès des points de vente ayant reçus des recommandations ne pourront pas être effectué.

Notions à évaluer	Modules	Techniques d'évaluation
Homogénéisation des données	MoPRD	Métrique Page Rank de Google
Amélioration de la prédiction	MoPRD et MoP	Méthode du score F1 de Herlocker et al. (2004)
Amélioration du démarrage à froid	MoPRD et MoP	Méthode du score F1 de Herlocker et al. (2004)
Pertinence de la classification des items recommandés	MoCI	Méthode du score F1 de Herlocker et al. (2004) et l'avis de l'expert
Automatisation partielle de l'analyse de l'expertise humaine	SRSE	Méthode du score F1 de Herlocker et al. (2004) et l'avis de l'expert

TABLE 4.3 – Méthodes d'évaluation du SRSE

4.3 Résultats des évaluations

Dans cette partie, nous allons présenter les résultats de nos évaluations suivant le tableau des méthodes d'évaluations de notre SRSE, énoncé précédemment.

4.3.1 Automatisation partielle de la démarche d'analyse de l'expert

L'automatisation partielle de la démarche d'analyse de l'expert est la contribution principale de cette thèse. L'objectif de notre SRSE est de numériser le traitement d'analyse de l'expert et de lui recommander des plans d'action qu'il va proposer aux points

de vente. De ce fait les résultats de l'évaluation de cette contribution seront obtenus à la fin du traitement du SRSE c'est-à-dire après la classification des items et la méthode d'évaluation qui est utilisée ici est l'avis de l'expert sur les recommandations produites et classifiées par notre SRSE. Les résultats concernant cette contribution seront abordés avec celle présentant la pertinence de la classification des items recommandés.

4.3.2 Homogénéisation des données

Dans notre évaluation, nous avons sélectionné quelques concepts, issus de notre ontologie et à partir de la métrique de Pr(P), nous avons cherché le nombre de relations que ces concepts possèdent en fonction des données sélectionnées. La Figure 4.8 présenté ci-dessous montre l'évolution du nombre de relations de nos différents concepts à chaque collecte de données.

Les concepts que nous avons utilisés pour valider notre hypothèse sur l'homogénéisation avec les outils du web sémantique sont obtenus à partir des données collectées et de leurs représentations au niveau de notre ontologie de gestion de compétences.

Pour réaliser cette évaluation, nous avons développé un programme de « Page Rank » en Python, utilisant l'API de Tensorflow. Nos données d'entrée sont représentées sous forme de matrices dans un fichier. Ces matrices sont formées à partir de la représentation des connaissances . Pour obtenir ces résultats nous avons chargé 90 questionnaires renseignés par les clients mystères pour chaque vague d'enquêtes. Ces questionnaires sont identiques pour toutes les vagues et sont modélisés sous la base des mêmes compétences à évaluer.

Pour évaluer notre hypothèse, nous allons comparer les résultats obtenus avec un autre chargement des données issues des enquêtes sans réaliser d'extraction et de représentation des connaissances de ces dernières. Ainsi, dans ce deuxième chargement nous n'allons pas utiliser notre algorithme d'extraction de connaissances ni l'ontologie. Les matrices sont ici construites à partir des données brutes issues des questionnaires. Cette expérimentation permet de vérifier notre hypothèse sur l'homogénéisation des données par les technologies du web sémantique. Le programme que nous avons écrit, basé sur la démarche de « Page Rank » est l'algorithme de Page Rank présenté dans l'annexe A.

La Figure 4.8 présente le résultat du premier chargement des données avec l'utilisation

de notre algorithme d'extraction et de représentation des connaissances par nos ontologies.

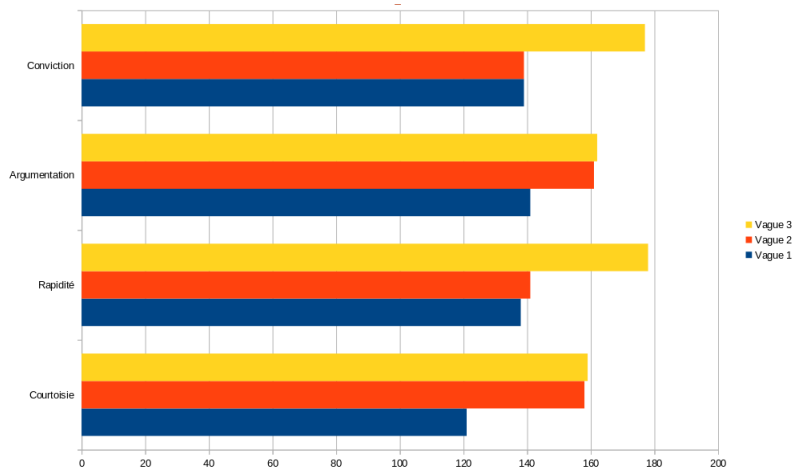

FIGURE 4.8 – Recherche de relations entre les concepts avec l'utilisation des technologies sémantiques pour l'homogénéisation.

La Figure 4.9 montre les résultats obtenus sans l'utilisation des technologies sémantiques sur les données sélectionnées des questionnaires d'enquêtes mystères.

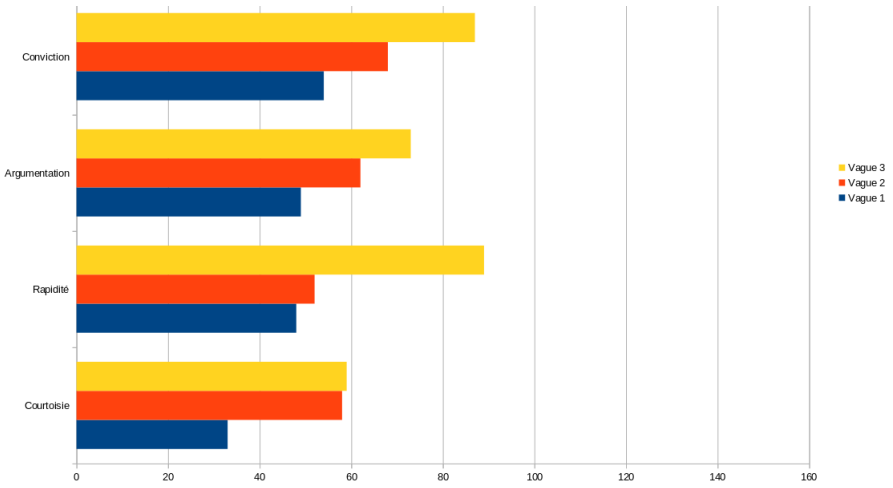

FIGURE 4.9 – Recherche de relations entre les concepts sans l'utilisation des technologies sémantiques.

Les résultats obtenus par cette évaluation permettent de valider l'hypothèse que les technologies du Web sémantique peuvent homogénéiser des données hétérogènes issues des résultats d'enquêtes. Nous avons en effet constaté en utilisant la métrique de page Rank, que le nombre de relations entre les différents concepts augmentent significativement en fonction du cumule des questionnaires des trois vagues. Cela signifie que les connaissances extraites des données d'enquêtes mystères sélectionnées ont été bien catégorisées et que les concepts qui permettent de les définir sont bien reliés. Si on prend par exemple le concept « Rapidité » , il est passé de 133 à 167 relations au cours des trois vagues d'enquête. Alors que pour le même concept dont la matrice est construite sur la base des données non appliquée aux technologies sémantiques, est passé de 47 à 87 relations. Cela signifie qu'à mesure que le nombre de questionnaires renseignés par les clients mystère augmente, la matrice qui décrit le concept s'est enrichie de relations sur la base des technologies sémantiques. De ce fait, l'utilisation de l'ontologie pour représenter les connaissances, permet d'homogénéiser les données collectées.

Dans un processus de recommandation la prédiction est incontournable. Ainsi pour obtenir une prédiction efficace, nous allons proposer une technique pour améliorer davantage la prédiction faite par les SR traditionnels en faisant appel aux technologies de l'apprentissage automatique.

4.3.3 Amélioration de la prédiction

Nous allons montrer ici que les technologies de l'apprentissage automatique peuvent améliorer la prédiction pour la recommandation. De ce fait pour réaliser nos prédictions, notre SRSE propose un modèle de recommandation basé sur le FColl comportemental (Esslimani & Igalens, 2008). Ce modèle exploite les observations relatives au comportement sur la base de données issues des données d'enquêtes mystères. Ces observations sont réalisées après la sélection et la prise en compte des précédentes recommandations qui ont été proposées.

Pour valider notre hypothèse, nous avons utilisé les données obtenues via les questionnaires des trois vagues d'enquêtes mystères. Dans notre cadre d'évaluation, les compétences utilisées précédemment sont reprises, car ces dernières représentent les items à recom-

mander. Les algorithmes 3 et 4, présentés dans l'annexe A sont utilisés pour la phase de prédiction.

Nous comparons ensuite nos résultats avec les résultats de l'algorithme de la recommandation basée sur la popularité sociale qui fait partie des techniques de recommandations les plus efficaces en termes de pertinence sur la prédiction. C'est une technique de recommandation basée sur le FColl proposé par Barman et Dabeer (2010). Les systèmes de ce type recommandent les items les plus populaires chez les amis de l'utilisateur courant.

Nous avons développé une application Python qui calcule la précision des meilleures K items pour chaque point de vente. Cette application utilise un ensemble de données obtenues des trois vagues d'enquêtes mystères. Tel qu'il est illustré sur la Figure 4.8 ci-dessous, l'évaluation est composée de trois phases. Notre programme python divise d'abord les data-sets dans des ensembles d'observations (40%), d'apprentissages (40 %) et des ensembles de tests (20%). Dans une deuxième phase, le programme utilise l'ensemble des apprentissages et des observations pour calculer les meilleures K prédictions pour chaque utilisateur en utilisant les algorithmes 3 et 4, présentés dans l'annexe A. Enfin, l'application utilise ces prédictions et l'ensemble de test pour calculer les scores F1. Le score F1 de Herlocker et al. (2004) est utilisé dans cette expérience pour mesurer la pertinence de chaque algorithme. Le score F1 varie entre 0 et 1. Nous calculons le F1 pour chaque point de vente ensuite nous calculons la valeur moyenne de ces scores F1. Le score F1 a été calculé pour les K recommandations de chaque point de vente.

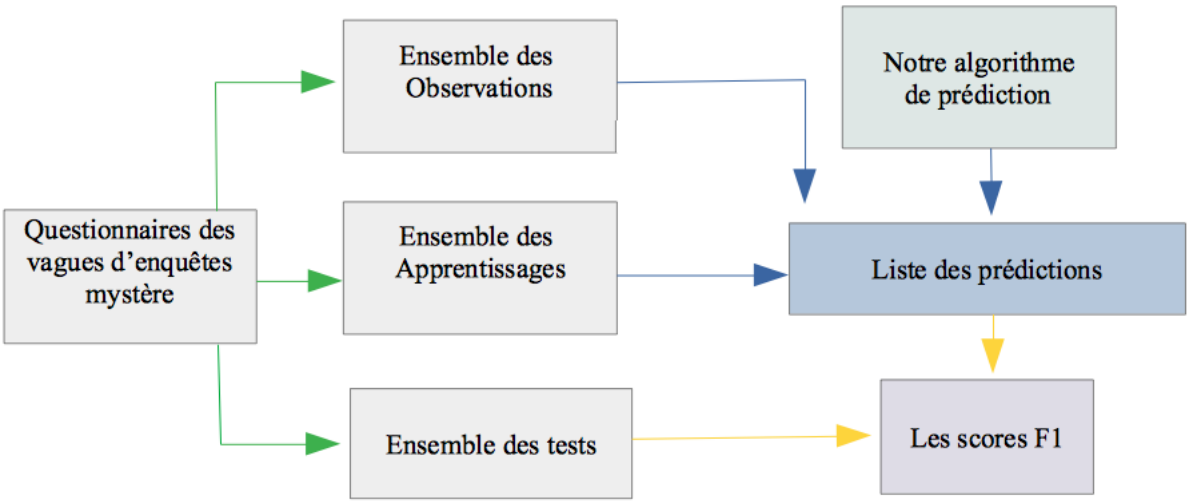

FIGURE 4.10 – Processus d'amélioration de la prédiction

L'ensemble des données extraites des questionnaires des trois vagues d'enquêtes mystères contient 207 points de vente. Les points de vente ont généré 507 évaluations de la part des clients mystère. Les résultats de l'analyse sont présentés dans le tableau 4.3. Ce tableau montre le score F1 moyen obtenu par notre algorithme pour différentes valeurs de K. Les valeurs en surbrillance indiquent le score F1 le plus performant. Ce tableau illustre également le pourcentage d'amélioration de notre algorithme par rapport au meilleur résultat des quatre autres algorithmes. Pour K = 1, notre algorithme surpasse l'algorithme de la popularité sociale de référence. Lorsque K varie de 1 à 7, notre algorithme dépasse aussi les autres et la pertinence des prédictions est améliorée de 0,03 à 0,80 par rapport à l'algorithme de popularité sociale qui occupe la deuxième place. Pour K = 10, l'algorithme proposé et la méthode de popularité des items ont la même pertinence et ont le meilleur score F1 par rapport aux deux autres algorithmes.

	Filtrage collaboratif basé sur l'utilisateur	**Filtrage basé sur l'item**	**Prédiction de l'algorithme de la popularité sociale**	**Les algorithmes 3 et 4**	**Amélioration**
Score F1 N1	0,07	0,12	0,21	0,27	0,06
Score F1 N2	0,09	0,17	0,24	0,30	0,06
Score F1 N3	0,09	0,17	0,30	0,41	0,11
Score F1 N4	0,12	0,15	0,34	0,41	0,70
Score F1 N5	0,11	0,13	0,34	0,42	0,80
Score F1 N6	0,10	0,16	0,37	0,42	0,05
Score F1 N7	0,15	0,18	0,42	0,45	0,03
Score F1 N8	0,13	0,18	0,47	0,47	0,00
Score F1 N9	0,12	0,17	0,47	0,47	0,00
Score F1 N10	0,19	0,18	0,47	0,47	0,00

TABLE 4.4 – Résultats du calcul du score F-1 des prédictions obtenues

4.3.4 Amélioration du démarrage à froid et formation de communautés de points de vente

Le démarrage a froid pour un nouveau point de vente est une problématique très importante (voir chapitre 1, page 50). Dans cette thèse nous estimons incontournable une expérimentation pour l'amélioration de ce problème en utilisant les technologies sémantiques et celles de l'apprentissage automatique. Dans le cadre de cette expérimentation, nous avons comparé les résultats obtenus par notre système et ceux proposés par un SR traditionnel. Tel qu'il est illustré sur la figure 4.9 ci-dessous, l'expérimentation est composée de trois phases. Les données passives et actives d'un nouvel utilisateur vont être collectées dans un premier temps. Pour rappel dans notre cadre d'étude l'utilisateur représente un point de vente. Ensuite, les connaissances seront extraites de ces données de cette organisation et à l'aide de la similarité sémantique, elles seront affectées à une communauté d'utilisateurs. En dernière phase, l'algorithme du plus proche voisin est utilisé sur la base de profils utilisateurs de la communauté à laquelle il a été associé et celui du nouveau point de vente. Ainsi, le nouveau point de vente se verra associer l'historique des recommandations que son plus proche voisin a déjà reçues.

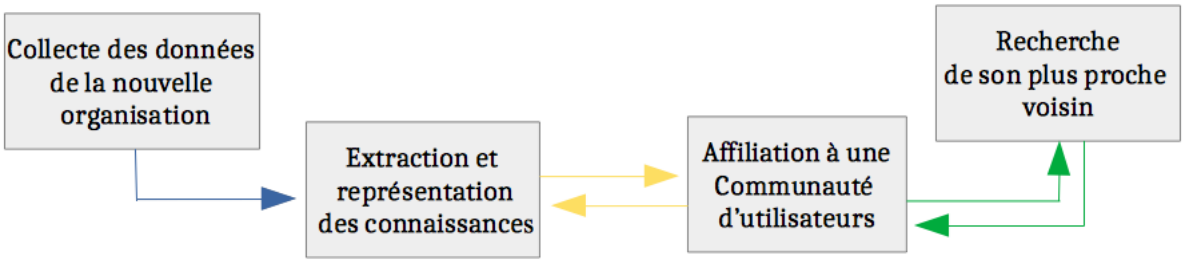

FIGURE 4.11 – Processus d'amélioration du démarrage à froid

Dès l'introduction du nouveau point de vente dans notre système, nous avons calculé la pertinence de l'ensemble des recommandations pour chaque point de vente. Ainsi on a la possibilité de voir la qualité des premiers items recommandés du nouveau point de vente. Ensuite, nous avons refait la même expérience sans notre technique d'amélioration avec un autre nouveau point de vente dans le but d'obtenir une comparaison des résultats et une validation de notre hypothèse d'amélioration du démarrage à froid.

Le graphe ci-dessous montre les résultats du calcul de pertinence des recommandations émises pour 11 points de vente qui partage la même communauté d'utilisateurs. Parmi ces 11 points de vente nous avons un nouveau point de vente numéro 7 à qui a été appliqué sous notre technique d'amélioration du démarrage froid. Les recommandations qui ont été faites sont des compétences sur la courtoisie, la rapidité, l'argumentation et sur la conviction. Pour une première recommandation du point de vente numéro 7 nous avons trouvé son « plus proche voisin », qui le point de vente numéro 5. Ensuite, un autre point de vente numéro 8 est introduit dans notre système, mais sans lui appliquer notre technique d'amélioration.

Les résultats de ces deux expériences montrent que par rapport au point de vente numéro 8, le point de vente numéro 7 reçoit des recommandations plus pertinentes. D'après les résultats obtenus on peut dire que notre proposition donne un bon résultat même s'il peut être encore amélioré.

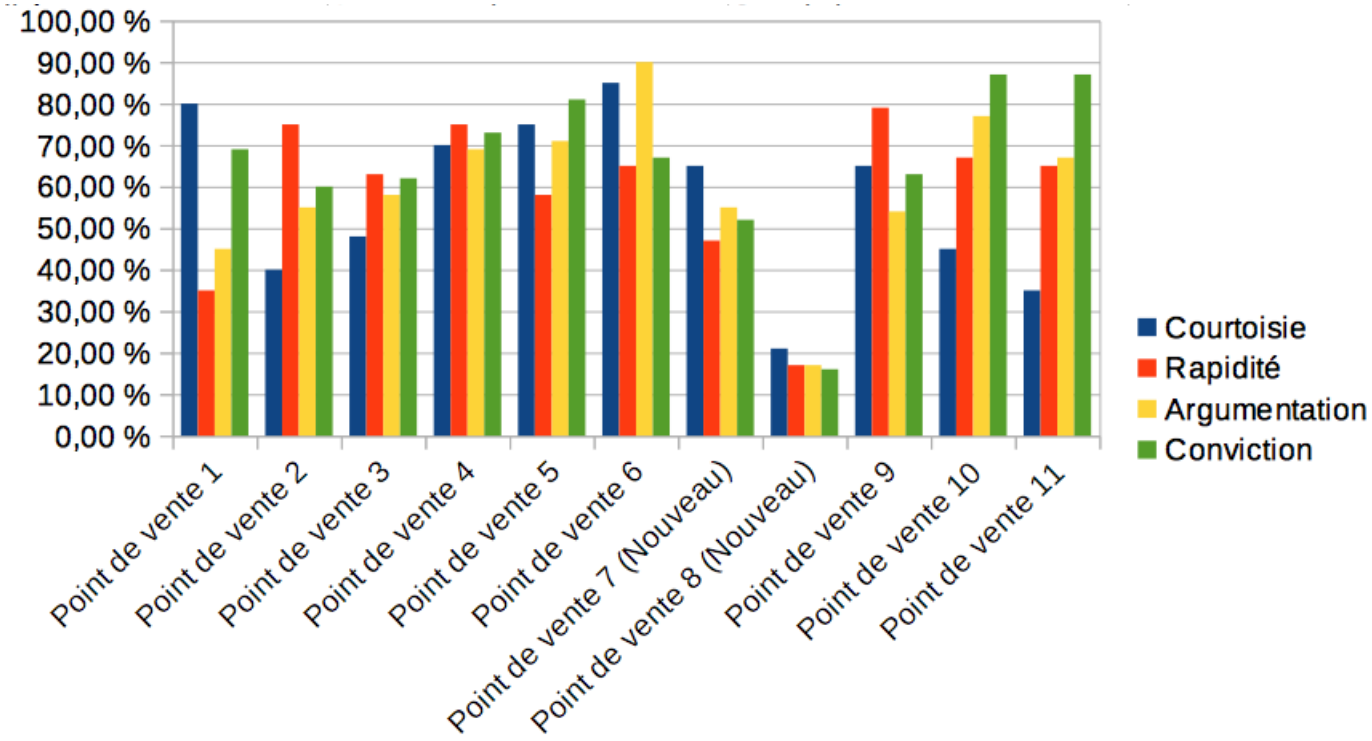

FIGURE 4.12 – Graphe de pertinence des recommandations

4.3.5 Classification et pertinence des items recommandés

Cette classification est réalisée sur la base de la pertinence des items recommandés après la phase de prédiction. Nous avons utilisé l'apprentissage supervisé avec l'algorithme de K-NN en prenant en compte l'ensemble des observations et données d'entrainement. Une partie de l'implémentation est présentée dans l'algorithme 5 présenté dans l'annexe A et en amont annoncée dans le chapitre 3.

Pour évaluer notre proposition nous avons réalisé une comparaison entre les besoins définis par le commanditaire des enquêtes mystères en début de traitement et les résultats par la méthode du score F1 de Herlocker et al. (2004) obtenus sur la pertinence des items recommandés dans la liste primaire, après la phase de prédiction. Ces besoins ont été établis par les commanditaires des enquêtes mystère avec des niveaux d'importances pour chacun de ces derniers.

Sur la base des observations et des données d'entrainement, le modèle d'apprentissage est mise à jour en prenant en compte les besoins qui ont été définis.

Le graphe ci-dessous montre la comparaison entre les pertinences des items et les niveaux d'importance des besoins de chaque item.

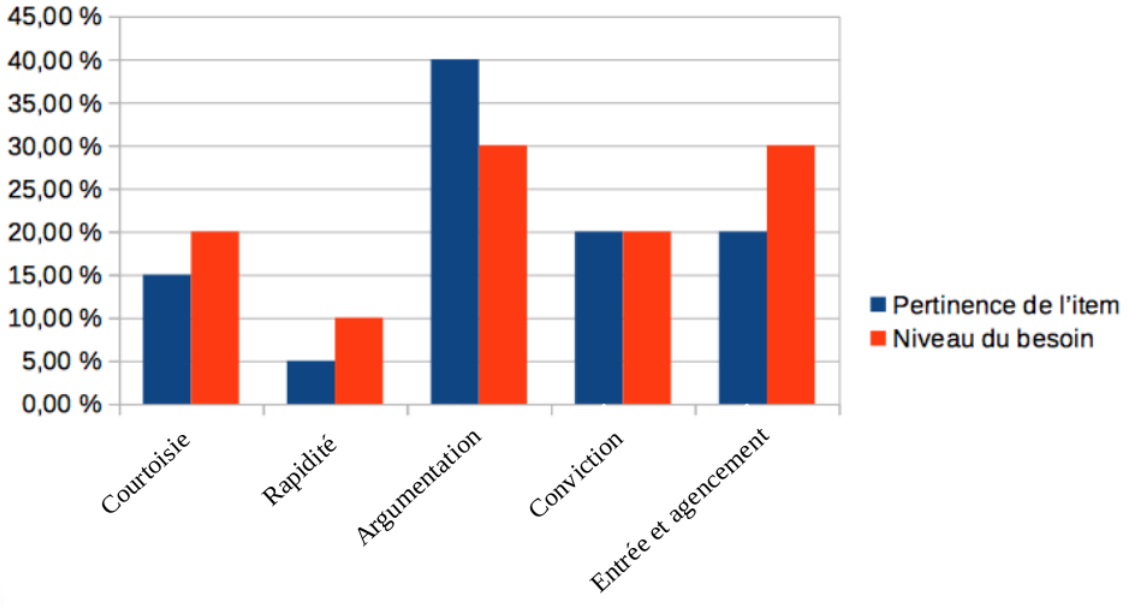

FIGURE 4.13 – Graphe de pertinence des recommandations vs les besoins fixés par l'organisation

Notre algorithme de classification nous permet de proposer à l'expert, la grille des items à recommander à l'organisation ci-dessous.

Rang	**Item**	**Nombre de récurrences**	**Taux de pertinence**	**Degrés d'importance**
1	Argumentation	97	40%	25 %
2	Entrée et agencement	45	20%	30 %
3	Conviction	25	20%	20 %
4	Courtoisie	15	15%	20 %
5	Rapidité	7	5%	5 %

TABLE 4.5 – Résultats calcul de pertinence des prédictions

Cette grille de lecture proposée à l'expert nous permet d'obtenir son évaluation par rapport aux items recommandés et à la classification de ces derniers. Le but de son évaluation est de voir si son raisonnement converge vers le nôtre et même temps de valider notre hypothèse sur l'automatisation partielle de l'expertise humaine. Pour valider notre proposition, 11 grilles de lectures destinées aux 11 points de vente ont été présentées à un expert pour sa prise décision finale. De ce fait, nous avons obtenu les résultats ci-dessous. Ces résultats sont les notes qui ont été émises par l'expert sur la cohérence des recommandations proposées dans les grilles de lecture pour les 11 points de vente. Les notations de l'expert porte sur la pertinence et le classement des items recommandés dans les grilles de lecture.

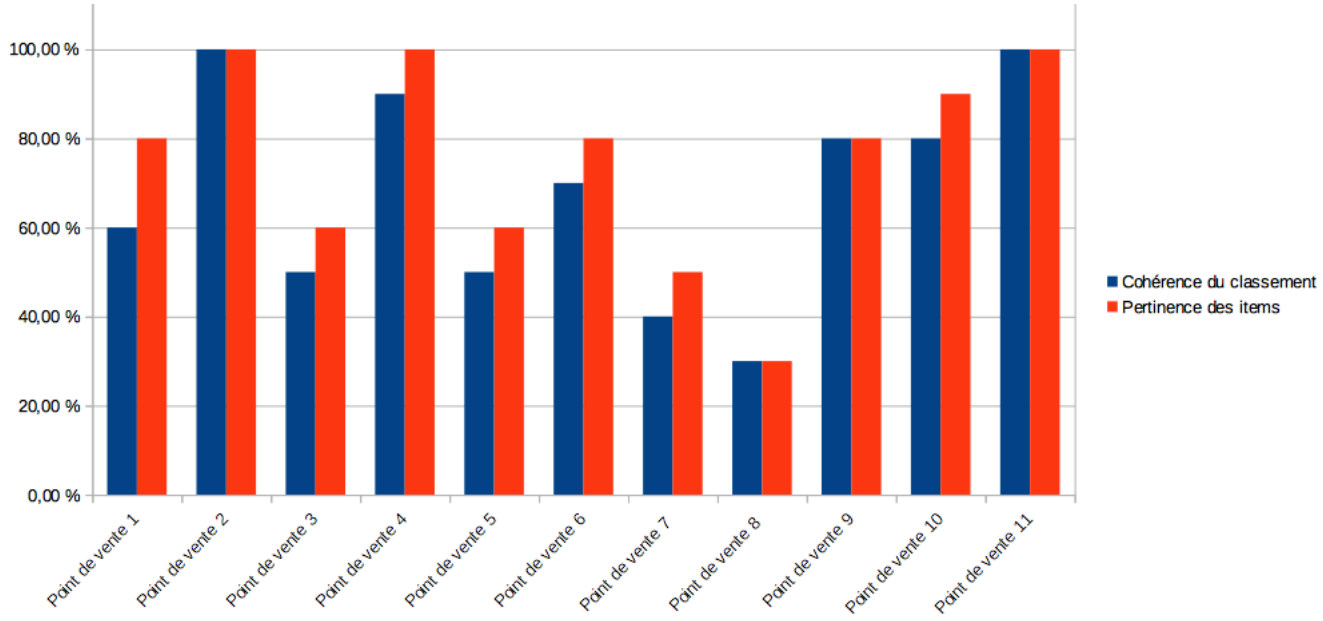

FIGURE 4.14 – Résultats fournis par l'expert sur la base des grilles de recommandation

Les résultats montrent que l'expert est en accord à 100% sur la cohérence classement et la pertinence des items avec deux points de vente. La grille de recommandation la moins convaincante aux yeux de l'expert est à 30% sur la cohérence du classement et sur la pertinence des items proposés. Ce qui est un résultat assez satisfaisant pour un système dont la durée d'apprentissage n'est pas longue. Nous avons remarqué que les points de vente 7 et 8 ayant reçu les notes les plus basses sont ceux qui ont été introduit récemment dans le système.

4.4 Discussions

Au cours de ces évaluations, nous avons étudié le fonctionnement et la pertinence de plusieurs fonctionnalités de notre SRSE. Quelques conclusions générales se dégagent de nos résultats, ainsi que plusieurs améliorations à apporter à notre SRSE.
Le SRSE que nous avons proposé dans cette thèse est une hybridation de deux technologies et de deux méthodes de recommandation. Ces deux technologies ont été utilisées ici pour soutenir ou encore renforcer notre algorithme de recommandation traditionnelle. La sémantique est utilisée pour donner du sens aux différentes données sélectionnées en faisant une extraction et une représentation des connaissances. Les technologies de l'intelligence artificielle, notamment l'apprentissage automatique viennent pour enrichir et d'optimiser les algorithmes de recommandation que nous proposons dans cette thèse.
Dans cette discussion nous allons situer notre proposition de SR par rapport à la littérature, ensuite présenter les limites de notre SR enfin résumer les différents apports que notre système propose.

Les résultats que nous avons obtenus des évaluations ont permis de justifier les hypothèses que nous avions annoncées dans l'introduction du manuscrit. En effet, dans nos hypothèses de recherche, nous avons noté que :

— le processus d'analyse des résultats d'enquêtes mystères peut être automatisé de manière partielle ;
— le problème de démarrage à froid pour un nouvel utilisateur peut être amélioré en utilisant les technologies du Web Sémantique ;
— les verrous liés à l'hétérogénéité des données sélectionnées peuvent être levés en faisant appel aux technologies du web sémantique ;
— l'amélioration de la prédiction pour la recommandation.

Nos résultats rejoignent ceux de Kunaver et Požrl (2017) ; Resnik (1995) ; Schafer et al. (2007), ces derniers ont travaillé sur les systèmes de recommandation basée sur la connaissance pour des données d'enquête, leurs champs de recherches sont proches à ceux présentés dans cette thèse. Le plus que nous apportons par rapport aux travaux de Kunaver et Požrl (2017) ; Resnik (1995) ; Schafer et al. (2007) est l'amélioration du

démarrage pour un nouvel utilisateur et l'hybridation entre les technologies de l'apprentissage automatique et celles du Web sémantique pour l'amélioration de la prédiction. Nos travaux sur l'amélioration du démarrage rejoignent ceux de Burke (2002) ; Schein et al. (2002) ; Safoury et Salah (2013) ; Berrichi et Djouaher (2020), contrairement à eux pour traiter ce problème, nous avons exploité les atouts de la représentation des connaissances par outils du Web sémantique et les technologies de l'intelligence artificielle, notamment l'apprentissage automatique. Néanmoins, durant nos expérimentations, nous avons fait face à des données dites complexes. Ces données complexes sont ici celles qui sont difficiles à inclure dans le processus de recommandation. Parmi ces données on peut citer : les données dont le lien avec le reste des données est difficilement détectables, ou encore difficiles à catégoriser.

Le comportement de notre système face aux données complexes est assez encourageant. En effet, en intégrant dans notre processus de recommandation, les technologies récentes comme la sémantique et celles de l'apprentissage automatique, on a obtenu une amélioration sur les limites suivantes :

— l'incohérence des données entre elles ;
— le démarrage à froid pour un nouvel utilisateur et un nouvel item ;
— la prise en compte de données « inutiles » ;
— la détection des items à évaluer ;
— la redondance des items recommandés ;

Limites de notre système

Le SRSE que nous proposons est performant en terme de pertinence, les résultats présentés l'ont confirmé. Ces résultats ont permis de valider nos différentes hypothèses. Par contre, notre SRSE comporte quelques limites :

1. un délai de traitement important durant les premiers lancements de l'apprentissage. On a pu atténuer ce problème en segmentant les différents data-sets que le SRSE reçois en entrée, mais ce qui reste quand même une problématique non négligeable ;
2. le démarrage à froid pour un nouvel utilisateur a été amélioré avec les technologies de l'apprentissage automatique et celles de la sémantique, mais le problème persiste. En

effet, le résultat obtenu sur la pertinence des items recommandés est encourageant, mais pas assez suffisant pour une première recommandation à un nouveau point de vente.

En guise de résumé, les apports présentés et évalués dans cette thèse ont été justifiées. Ces derniers ont permis de valider nos différentes hypothèses de recherche qui portant sur :

— l'automatisation partielle de l'expertise humaine en proposant un système de recommandation sémantique enrichi ;
— la résolution du verrou d'hétérogénéité des données par les technologies du web sémantique ;
— l'amélioration du démarrage à froid, dans le cas d'un nouvel utilisateur dans le système ;
— l'amélioration de la pertinence des prédictions pour la recommandation ;
— la réalisation une bonne classification des items recommandés.

Néanmoins, le problème de démarrage à froid pour un nouvel utilisateur peut être amélioré.

*

* *

Dans ce dernier chapitre de notre manuscrit de thèse, l'implémentation et les évaluations de notre proposition ont été présentées. Dans un premier temps, les technologies utilisées pour mettre en place notre prototype sont présentées. Ensuite, le fonctionnement de notre système a été présenté avec les différentes étapes de traitement. Enfin, nous avons présenté les résultats obtenus durant nos différentes expérimentations et une discussion a été effectuée autour des résultats obtenus et par rapport à la littérature existante.
Nous avons montré que les technologies de l'apprentissage automatique ont enrichie notre SRSE. En effet, ces technologies issues de l'intelligence artificielle ont été utiles dans les étapes suivantes : la sélection de données ; la formation de communautés d'utilisateurs ; la prédiction des items à recommander ; la classification des items recommandés.
Les technologies d'apprentissage automatique ont été très efficaces dans notre processus de recommandation, car nous permettant de baser nos traitements sur des modèles qui s'améliorent davantage. La sémantique a aussi joué un rôle très important dans notre pro-

cessus de recommandation, notamment pour l'extraction, la représentation des connaissances et la mise en relation des données entre elles.

« Toute connaissance acquise sur la connaissance devient un moyen de connaissance éclairant la connaissance qui a permis de l'acquérir. »
(Morin, 1992, p. 232)

Conclusion et perspectives

Depuis quelques années de nombreuses applications informatiques de collecte de données ont été développées et l'expression « Big Data » est maintenant utilisée dans le monde numérique pour signifier la production de quantités massives de données. Face à cette grande quantité de données, de nouveaux outils informatiques ont été développés pour les analyser. L'objectif de tels outils est de proposer à utilisateur un système informatique lui permettant de faire des choix, un système permettant de filtrer l'information et de l'adapter au profil ou à la recherche de l'utilisateur pour recommander des informations utiles pour ce dernier. Ces systèmes de recommandation sont pilotés par des algorithmes informatiques et accumulent une très grande quantité de données sur les utilisateurs, en les croisant avec d'autres données (comportements des utilisateurs similaires, des produits), seraient ainsi en mesure de prédire les produits qui seront utiles pour cet utilisateur.

La recherche présentée dans ce manuscrit de thèse porte sur l'aide à la prise de décision, notamment sur la proposition d'un système de recommandation sémantique enrichi en application au domaine du e-marketing. Le système que nous proposons permet de faire face à notre problématique de recherche liée à l'exploitation d'une masse de données issue des enquêtes mystères. Ces enquêtes mystères constituent en marketing un moyen pour vérifier concrètement la bonne commercialisation des produits d'un ou plusieurs points de vente et sont réalisées par de « faux » clients recrutés et missionnés par les entreprises pour évaluer leur réseau de ventes. Ces derniers ont pour rôle de simuler un processus d'achat basé sur un scénario prédéfini par l'entreprise sollicitant l'enquête mystère. Les données produites à travers ces enquêtes mystères sont très complexes pour les experts et ces derniers doivent passer beaucoup de temps pour trouver les points à améliorer au sein du réseau de vente qui a sollicité les enquêtes mystères. Notre recherche a été

appliquée sur le logiciel Retaily, développé par l'entreprise Effet B spécialisée dans le développement informatique. L'objectif de cette thèse est de proposer un système permettant de guider l'expert dans son analyse des données d'enquêtes mystères et dans sa prise de décision sur les recommandations d'améliorations à proposer aux différents points de vente. Notre hypothèse de recherche principale est que l'expertise humaine peut être automatisée de manière partielle en utilisant les technologies du web sémantique et de l'apprentissage automatique. Les technologies de l'apprentissage automatique viennent enrichir notre proposition de système de recommandation proposé en début de thèse. En effet, en première année de doctorat, nous avions orienté nos recherches sur l'exploitation des technologies du web sémantique pour évaluer l'opportunité d'extraire, décrire et de modéliser l'expertise d'un agent humain en vue de l'implémenter dans un système informatique capable d'interpréter des données issues d'enquêtes mystères. Avec les résultats satisfaisants des recherches menées ces dernières années sur les technologies d'apprentissage automatique, présentés dans le chapitre 1 (pages 24-33), précisément sur les systèmes de recommandation, nous avons décidé d'inclure l'apprentissage automatique dans notre proposition pour anticiper la valorisation ou la préférence qu'un client mystère attribuerait par exemple à un ensemble d'aspects commerciaux au sein d'un point de vente. En effet, nos recherches sont appliquées dans les champs du marketing, particulièrement dans le domaine de la commercialisation de produits en magasin.

Ce manuscrit de thèse est composé de quatre chapitres, deux chapitres présentant l'état de l'art sur les systèmes de recommandation et sur l'expertise humaine et la représentation des connaissances. Les deux de derniers chapitres portent sur nos différentes contributions apportées dans cette thèse ; sur l'implémentation de nos propositions et sur l'évaluation de notre système de recommandation sémantique enrichi. Après avoir présenté les principales caractéristiques des systèmes de recommandation et réalisé un état de l'art, les limites de ces derniers ont été exposées. Nous nous sommes focalisés sur la problématique liée au démarrage à froid. Comme présenté dans le chapitre 1(page 50), le démarrage à froid survient lorsque les recommandations sont nécessaires pour des utilisateurs (ou objets) pour lesquels nous ne possédons aucune information. Nous avons vu dans ce chapitre 1 qu'il y a plusieurs problèmes liés au démarrage à froid :

nouvel utilisateur, nouveau produit et apprentissage (Berrichi & Djouaher, 2020). Nous avons travaillé dans cette thèse sur l'amélioration de la problématique liée au démarrage à froid dans le cas d'un nouvel utilisateur, car ce problème peut impacter de manière significative le système que nous proposons. De ce fait, dans cette thèse une contribution a été faite dans l'objectif d'améliorer les limites liées au démarrage à froid. En effet, pour faire face, nous avons proposé une méthode basée sur les technologies sémantiques, précisément pour la représentation sémantique des données et sur la recherche de similarité sémantique entre profil des points de vente. Nous avons une méthode permettant d'améliorer cette problématique. La méthode que nous avons proposée et implémentée est basée sur les technologies du Web Sémantique. Les résultats obtenus après évaluation sont très satisfaisants. Pour justifier l'amélioration du problème du démarrage à froid, nous avons comparé les résultats de notre méthode d'amélioration et les résultats sans l'utilisation de la méthode que nous proposons.

Les technologies du web sémantique, particulièrement les ontologies et les outils d'extraction de connaissances ont occupé une place importante dans le système de recommandation sémantique enrichi. En plus d'être utilisées dans l'amélioration du démarrage à froid, elles ont été appliquées aux données sélectionnées à partir des résultats d'enquêtes mystères produits par Retaily pour lever les verrous liés à l'hétérogénéité des données sélectionnées. En effet, l'homogénéisation des données sélectionnées par l'extraction et la représentation des connaissances font partie des défis qui ont été fixés dans cette thèse. Cette contribution d'homogénéisation, nous a conduit à extraire et à décrire les connaissances à partir des données sélectionnées, mais aussi à préparer et mieux structurer les connaissances avec les relations qui les relient.

Après la préparation, l'extraction et la représentation des connaissances, nous avons proposé une méthode basée sur l'apprentissage automatique et est utilisée pour une phase de prédiction pour générer une première liste de recommandations. Ces recommandations sont des plans d'action visant l'amélioration des forces de vente des différents points de vente. La méthode de prédiction que nous proposons est une amélioration de la méthode classique des systèmes de recommandation par FColl. En effet, notre proces-

sus de prédiction utilise les algorithmes d'apprentissage automatique et celui du FColl. Pour justifier notre hypothèse d'amélioration de la prédiction, nous avons comparé les résultats obtenu par la méthode que nous proposons et ceux résultants de la méthode de recommandation basée sur la popularité sociale qui est le SR le plus performant en termes de prédiction, selon (Barman & Dabeer, 2010).

Pour mieux guider l'expert dans sa prise de décision finale, nous avons proposé une méthode de classification des plans d'actions à recommander à l'issue de la phase de prédiction permettant de signaler à l'expert les plans d'action les plus pertinents au moins apte à apporter une amélioration significative aux points de vente. Cette classification est notre dernière contribution dans cette thèse basée sur un algorithme d'apprentissage automatique. Elle représente la dernière étape de notre de système recommandation et propose à l'expert les recommandations obtenues dans une grille de lecture que nous avons modélisée.

Pour évaluer et valider nos contributions, ces méthodes ont été modélisées et implémentées. Ces modélisations nous ont conduits à proposer une architecture système composée de trois modules.

Pour nos perspectives de recherche, nous allons explorer d'autres domaines d'application pour justifier interopérabilité du système que nous proposons dans cette thèse. Durant cette thèse, nous avons déjà eu à travailler sur quelques axes liés l'apprentissage en ligne et aux technologies sémantiques. Ainsi nous avons appliqué à notre système de recommandation sémantique le traitement pour la proposition de ressource pédagogique sur la base du profile de l'apprenant. Ces perspectives de recherche au sein de l'entreprise EFFET B seront ancrées d'une part dans le domaine de la santé, d'autre part dans le domaine de formations en ligne et de la gestion des compétences de l'apprenant.

L'entreprise Effet B travail sur une application de gestion de la santé, cette application mobile est nommée « Mia Hypnose » et a pour objectif de guider ses utilisateurs dans le suivi de leurs traitements. Les recherches qui vont être menées sur l'application mobile

sont orientées sur la recommandation par rapport à la santé et à l'hygiène de vie. L'objectif de cette perspective de recherche est de centrer des informations médicales en fonction du profile du malade et ces informations peuvent consister en :

1. des conseils d'experts sur la façon de faire face à une maladie ;
2. des définitions de maladies en général qui aident à comprendre la terminologie médicale ;
3. des plans de soins qui pourraient empêcher les patients d'agir contre les règles suggérées par la médecine factuelle ;
4. des conseils pour une vie plus saine ou des informations sur l'alimentation.

Toujours dans les champs de la recherche d'information, notamment sur les systèmes de recommandation sémantique enrichi, l'entreprise Effet B travail sur la recommandation de compétences des apprenants à travers un livret numérique de l'alternance, Studea. La plateforme Studea est développé par Effet B et destinée à la gestion de la procédure d'alternance. La recherche qui est prévue sur ce livret numérique est orientée sur l'évaluation des compétences des apprenants et vers la recommandation de ressources pédagogiques (Mbaye, 2018). L'objectif de cette perspective de recherche est de proposer une approche basée sur l'utilisation des référentiels des compétences afin d'évaluer le profil d'un apprenant et par la suite de recommander à ce dernier et à son tuteur pédagogique des activités pertinentes pour obtenir ou améliorer les compétences nécessaires au cours de sa formation. Les référentiels de compétences sont un ensemble unique de niveaux de compétences en fonction de la formation pour les contrats d'apprentissage. Les référentiels de compétences permettent d'évaluer le niveau de l'apprenant tout au long de son cursus, Ces référentiels sont définis et publiés par France compétences [3]. Les principales questions de recherche sur cette étude sont :

1. Comment construire un modèle de référentiels des compétences génériques basées sur la sémantique par l'extraction des connaissances afin de vérifier et de valider les contraintes d'élaboration et de réutilisation de tout référentiel de compétences lié aux plateformes d'évaluations ?
2. Comment évaluer le profil d'un apprenant à partir d'un référentiel des compétences et à l'aide des traces d'activités ? Nous souhaitons proposer une méthode autour de

3. https ://www.francecompetences.fr

l'apprentissage automatique basée sur les techniques du clustering.

3. Comment proposer un système de recommandation intelligent capable d'identifier les besoins de l'apprenant et de le guider au cours de sa formation ? Le système que nous voulons mettre en place sera basé sur le profil de l'apprenant et sur l'état du référentiel en cours d'évaluation.

Liste des publications

ACTI

Mbaye, B. (2018). Semantic-Based Collaborative Filtering to Improve Visitor Cold Start in Recommender Systems. World Academy of Science, Engineering and Technology, International Science Index, Computer and Information Engineering, 12(3), 2562.

Mbaye, B. (2018). Ontology-based collaborative filtering recommendation for e-learning. In Proceedings of the 17th International Conference on Informatics in Economy (IE 2018) Education, Research Business Technologies, Iasi, Romania, ISSN-L 2247-1480, page 585-592.

Mbaye, B. (2018). Recommender System : Collaborative Filtering of e-Learning Resources. In Proceedings 12th International Conference on e-Learning, International Association for Development of the Information Society, 17 – 19 July 2018, Madrid, Spain, ,ISBN : 978-989-8533-78-4, page 213-217.

Mbaye, B. (2018). Representation of expert knowledge for e-learning. In Proceedings of The singapore education technology conference 2018, Singapore, Singapore, ISBN :978-981-11-6603-7, page 142-147.

Mbaye, B. (2019). Recommender system using unsupervised machine learning for satisfaction surveys . In Proceedings of 4th International Conference on Big Data Analytics, Data Mining and Computational Intelligence 16 – 18 July 2019, Porto, Portugal, ISBN : 978-989-8533-92-0 page 251-255.

Mbaye, B. (2019). Organisation of knowledge from traces of human learning. In Proceedings 13th International Conference on e-Learning, 17 – 19 July 2019, Porto, Portugal, ISBN : 978-989-8533- 88-3, page 361-365.

Mbaye, B. (2020). Collaborative filtering combined with machine learning for satis-

faction surveys. In International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering Vol.3, Iss.2, 6th World Machine Learning and Deep Learning Congress October 24-25, 2019 held at Helsinki, Finland.

C-COM

Mbaye, B. (2018). Système de recommandation sémantique par filtrage collaboratif : amélioration du démarrage A froid. In Journée Jeunes Chercheurs en SIC–, GERiiCO 12e édition, Lille, France, 30 mai 2018.

Mbaye, B. (2019). Système de recommandation sémantique. Application au domaine du e- marketing. In Journées doctorales 2019, Montbéliard, France, 03-04 Juillet 2019.

Références

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. In *Proceedings of the anti-phishing working groups 2nd annual ecrime researchers summit* (pp. 60–69).

Aciar, S., Zhang, D., Simoff, S., & Debenham, J. (2007). Informed recommender : Basing recommendations on consumer product reviews. *IEEE Intelligent systems*, *22*(3), 39–47.

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, *16*(1), 3–9.

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

Aiman-Smith, L., Scullen, S. E., & Barr, S. H. (2002). Conducting studies of decision making in organizational contexts : A tutorial for policy-capturing and other regression-based techniques. *Organizational Research Methods*, *5*(4), 388–414.

Akdag, H., & Khoukhi, F. (1994). Une approche logico-symbolique dans les systèmes experts. *ERGO-IA'94*.

Ali, M., Jung, L. T., Abdel-Aty, A.-H., Abubakar, M. Y., Elhoseny, M., & Ali, I. (2020). Semantic-k-nn algorithm : an enhanced version of traditional k-nn algorithm. *Expert Systems with Applications*, *151*, 113374.

Aljukhadar, M., Senecal, S., & Daoust, C.-E. (2012). Using recommendation agents to cope with information overload. *International Journal of Electronic Commerce*, *17*(2), 41–70.

Allard-Poesi, F., Maréchal, C., et al. (1999). Construction de l'objet de la recherche. *Méthodes de recherche en management*, *3*, 34–57.

Allard-Poesi, F., Perret, V., et al. (2014). Fondements épistémologiques de la recherche. *Méthodes de recherche en management*, 14–46.

Amatriain, X. (2013). Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, *14*(2), 37–48.

Amatriain, X., & Basilico, J. (2015). Recommender systems in industry : A netflix case study. In *Recommender systems handbook* (pp. 385–419). Springer.

Amudha, R., Nalini, R., Alamelu, R., Hemalatha, K., & Visvanaath, M. (2018). Impact of social media network in experiential mystery shopping. In *2018 international conference on computation of power, energy, information and communication (iccpeic)* (pp. 103–107).

Armstrong, D. M. (1978). *Nominalism and realism : Volume 1 : Universals and scientific realism* (Vol. 1). CUP Archive.

Armstrong, J. S. (1991). Prediction of consumer behavior by experts and novices. *Journal of Consumer Research*, *18*(2), 251–256.

Armstrong, J. S., Adya, M., & Collopy, F. (2001). Rule-based forecasting : Using judgment in time-series extrapolation. In *Principles of forecasting* (pp. 259–282). Springer.

Arregle, J.-L., Cauvin, E., Ghertman, M., et al. (2000). *Les— nouvelles approches de la gestion des organisations* (Rapport technique).

Aussenac-Gilles, N. (2005). *Méthodes ascendantes pour l'ingénierie des connaissances* (Thèse de doctorat non publiée).

Balabanović, M., & Shoham, Y. (1997). Fab : content-based, collaborative recommendation. *Communications of the ACM*, *40*(3), 66–72.

Baloian, N., Galdames, P., Collazos, C. A., & Guerrero, L. A. (2004). A model for a collaborative recommender system for multimedia learning material. In *International conference on collaboration and technology* (pp. 281–288).

Barkat, O. (2017). *Utilisation conjointe des ontologies et du contexte pour la conception des systèmes de stockage de données* (Thèse de doctorat non publiée). ISAE-ENSMA Ecole Nationale Supérieure de Mécanique et d'Aérotechique-Poitiers.

Barman, K., & Dabeer, O. (2010). Local popularity based collaborative filters. In *2010 ieee international symposium on information theory* (pp. 1668–1672).

Basu, C., Hirsh, H., Cohen, W., et al. (1998). Recommendation as classification : Using social and content-based information in recommendation. In *Aaai/iaai* (pp. 714–720).

Basu, C., Hirsh, H., Cohen, W. W., & Nevill-Manning, C. (2001). Technical paper recommendation : A study in combining multiple information sources. *Journal of Artificial Intelligence Research*, *14*, 231–252.

Batsell, R. R., & Lodish, L. M. (1981). A model and measurement methodology for predicting individual consumer choice. *Journal of Marketing Research*, *18*(1), 1–12.

Bawden, D., Holtham, C., & Courtney, N. (1999). Perspectives on information overload. In *Aslib proceedings.*

Beam, M. A., Hutchens, M. J., & Hmielowski, J. D. (2018). Facebook news and (de) polarization : reinforcing spirals in the 2016 us election. *Information, Communication & Society*, *21*(7), 940–958.

Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings of the third acm international conference on web search and data mining* (pp. 291–300).

Bell, R. M., & Koren, Y. (2007). Lessons from the netflix prize challenge. *SiGKDD Explorations*, *9*(2), 75–79.

Belloui, A. (2008). Lusage des concepts du web smantique dans le filtrage dinformation collaboratif. *Institut National dInformatique Alger*.

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for ai. *Communications of the ACM*, *64*(7), 58–65.

Bennett, J., Lanning, S., et al. (2007). The netflix prize. In *Proceedings of kdd cup and workshop* (Vol. 2007, p. 35).

Benouaret, I. (2017). *Un système de recommandation contextuel et composite pour la visite personnalisée de sites culturels* (Thèse de doctorat non publiée). Université de Technologie de Compiègne.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, *284*(5), 28–37.

Berners-Lee, T., et al. (1998). *Semantic web road map.*

Berrichi, T., & Djouaher, L. (2020). *Problème du démarrage à froid dans les systèmes de recommandation* (Thèse de doctorat non publiée). Université Mouloud Mammeri.

Bhavsar, H., & Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(4), 2231–2307.

Bidault, A. (2002). Affinement de requêtes posées à un médiateur. *Université Paris XI, Orsay, Paris, France*.

Biernacki, C. (1997). *Choix de modèles en classification* (Thèse de doctorat non publiée). Compiègne.

Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). Knn based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, *7*(1), 61–70.

Bisseret, A. (1995). *Représentation et décision experte : Psychologie cognitive de la décision chez les aiguilleurs du ciel*. Octares éditions.

Blandin, A., Lecorvé, G., & Battistelli, D. (2019). *Prédiction de recommandations d'âge pour l'accès à des enfants à des textes* (Thèse de doctorat non publiée). Univ Rennes, CNRS, IRISA, France.

Blattberg, R. C., & Hoch, S. J. (1991). Modèles à base de données et intuition managériale : 50% modèle+ 50% manager. *Recherche et Applications en Marketing (French Edition)*, *6*(4), 79–98.

Boose, J. H. (1986). Ets—a system for the transfer of human expertise. In *Knowledge based problem solving* (pp. 112–165).

Boose, J. H., Bradshaw, J. M., Kitto, C. M., & Shema, D. B. (1989). From ets to aquinas : Six years of knowledge acquisition tool development. In *Proceedings of ekaw-89 : Third european workshop on knowledge acquisition for knowledge-based systems* (pp. 502–516).

Bouchereau, A. (2020). *Les objets connectés au service de l'apprentissage* (Thèse de doctorat non publiée). Université Bourgogne Franche-Comté.

Bouchindhomme, C., & Rochlitz, R. (1992). „temps et récit" de paul ricœur en débat.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the fourteenth conference on*

uncertainty in artificial intelligence (pp. 43–52).

Brown, A. K., O'Connor, P. J., Roberts, T. E., Wakefield, R. J., Karim, Z., & Emery, P. (2005). Recommendations for musculoskeletal ultrasonography by rheumatologists : setting global standards for best practice by expert consensus. *Arthritis Care & Research*, *53*(1), 83–92.

Brown, P. J., Bovey, J. D., & Chen, X. (1997). Context-aware applications : from the laboratory to the marketplace. *IEEE personal communications*, *4*(5), 58–64.

Burke, R. (2002). Hybrid recommender systems : Survey and experiments. *User modeling and user-adapted interaction*, *12*(4), 331–370.

Burke, R. (2007). Hybrid web recommender systems. In *The adaptive web* (pp. 377–408). Springer.

Burton-Jones, A. (1999). The knowledge-based firm : strategies for growth and competitive advantage. *Training & Development in Australia*, *26*(6).

Bustillo, A., Pimenov, D. Y., Mia, M., & Kapłonek, W. (2021). Machine-learning for automatic prediction of flatness deviation considering the wear of the face mill teeth. *Journal of Intelligent Manufacturing*, *32*(3), 895–912.

Caelen, J., & Villaseñor, L. (1997). Dialogue homme-machine et apprentissage. *Apprentissage par l'interaction*, 83–117.

Calabretto, S., Roussey, C., & Harrathi, F. (2009). Recherche d'information sémantique multilingue. In *7ème colloque du chapitre français de l'isko. intelligence collective et organisation des connaissances.* (p. inconnue).

Carré, F., Brion, R., Douard, H., Marcadet, D., Leenhardt, A., Marçon, F., & Lusson, J. (2009). Recommandations concernant le contenu du bilan cardiovasculaire de la visite de non contre indication à la pratique du sport en compétition entre 12 et 35 ans. *Arch Mal Coeur*, *182*, 41–3.

Chakravarti, D., Mitchell, A., & Staelin, R. (1981). Judgment based marketing decision models : Problems and possible solutions. *Journal of Marketing*, *45*(4), 13–23.

Champagne, I. (2004). *Méthodes de factorisation des équations aux dérivées partielles.* (Thèse de doctorat non publiée). Ecole Polytechnique X.

Chandrasekaran, B. (1987). Towards a functional architecture for intelligence based on generic information processing tasks. In *Ijcai* (Vol. 87, pp. 1183–1192).

Chanier, T., & Ciekanski, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, *13*.

Chantrain, G. (2017). *Éléments de la terminologie du temps dans les textes égyptiens (de l'ancien empire à la troisième période intermédiaire) : étude d'un réseau sémantique en diachronie* (Thèse de doctorat non publiée). UCL-Université Catholique de Louvain.

Charlet, J. (2002). *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales* (Thèse de doctorat non publiée). Université Pierre et Marie Curie-Paris VI.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive psychology*, *17*(4), 391–416.

Cherif, W. (2018). Optimization of k-nn algorithm by clustering and reliability coefficients : application to breast-cancer diagnosis. *Procedia Computer Science*, *127*, 293–299.

Chi, M. T., Glaser, R., & Farr, M. J. (2014). *The nature of expertise*. Psychology Press.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combing content-based and collaborative filters in an online newspaper.

Cleuziou, G., Clavier, V., & Martin, L. (2003). Une méthode de regroupement de mots fondée sur la recherche de cliques dans un graphe de cooccurrences. *Proceedings of Rencontres Terminologie et Intelligence Artificielle, France*, 179–182.

Coenen-Huther, J. (2007). Classifications, typologies et rapport aux valeurs. *Revue européenne des sciences sociales. European Journal of Social Sciences*(XLV-138), 27–40.

Cord, M., & Cunningham, P. (2008). *Machine learning techniques for multimedia : case studies on organization and retrieval*. Springer Science & Business Media.

Costa, H., & Macedo, L. (2013). Emotion-based recommender system for overcoming the problem of information overload. In *International conference on practical applications of agents and multi-agent systems* (pp. 178–189).

Cottet, F. (2020). *Traitement des signaux et acquisition de données : cours et exercices*

corrigés. Dunod.

Crettez, J.-P., & Lorette, G. (1998). *Reconnaissance de l'écriture manuscrite*. Ed. Techniques Ingénieur.

Cron, W. L., & Sobol, M. G. (1983). The relationship between computerization and performance : a strategy for maximizing the economic benefits of computerization. *Information & management*, *6*(3), 171–181.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia* (pp. 21–49). Springer.

Dalkey, N. C. (1968a). *Experiments in group prediction* (Rapport technique). RAND CORP SANTA MONICA CALIF.

Dalkey, N. C. (1968b). *Predicting the future* (Rapport technique). RAND CORP SANTA MONICA CA.

Davenport, T. H., Prusak, L., et al. (1998). *Working knowledge : How organizations manage what they know*. Harvard Business Press.

Davis, N. (2011). Information overload, reloaded. *Bulletin of the American Society for Information Science and Technology*, *37*(5), 45–49.

Davis, R. (1979). Interactive transfer of expertise : Acquisition of new inference rules. *Artificial intelligence*, *12*(2), 121–157.

Décaudin, J.-M., Elayoubi, M., & IAE-Toulouse, I. (2009). Le concept d'expert : une définition dans le champ du marketing. *8ème congrès*.

De Gemmis, M., Lops, P., Musto, C., Narducci, F., & Semeraro, G. (2015). Semantics-aware content-based recommender systems. In *Recommender systems handbook* (pp. 119–159). Springer.

Dennis, C., Marsland, D., & Cockett, W. (2001). The mystery of consumer behaviour : market segmentation and shoppers' choices of shopping centres.

Denoyer, L., Baskiotis, N., & Schwandler, O. (2016). Apprentissage statistique.

Desclés, J.-P. (1987). Réseaux sémantiques : la nature logique et linguistique des relateurs. *Langages*(87), 55–78.

Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 143–177.

de Villaseñor, Y. F. O. (1989). *Gestion de connaissances pour des applications du domaine*

de la parole (Thèse de doctorat non publiée).

Dramé, K. (2014). *Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical* (Thèse de doctorat non publiée). Bordeaux.

Dudognon, D., Hubert, G., & Ralalason, B. (2010). Proxigénéa : Une mesure de similarité conceptuelle. In *Proceedings of the colloque veille strategique scientifique et technologique (vsst 2010)*.

Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). Sewep : using site semantics and a taxonomy to enhance the web personalization process. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 99–108).

Elavarasi, S. A., Akilandeswari, J., & Menaga, K. (2014). A survey on semantic similarity measure. *International Journal of Research in Advent Technology*, *2*(3), 389–398.

El Bouhissi, H., Bazizi, R., et al. (2020). *Construction d'ontologies à l'aide de significations unifiées et liées pour le web sémantique* (Thèse de doctorat non publiée). Univ. A/Mira Bejaia.

Esslimani, B., & Igalens, J. (2008). Rôle de l'empowerment dans le développement d'un comportement orienté client chez le personnel en contact avec la clientèle. *Revue de gestion des ressources humaines*(2), 17–29.

Ezz, M., & Elshenawy, A. (2020). Adaptive recommendation system using machine learning algorithms for predicting student's best academic program. *Education and Information Technologies*, *25*(4), 2733–2746.

Farzan, R., & Brusilovsky, P. (2011). Encouraging user participation in a course recommender system : An impact on user behavior. *Computers in Human Behavior*, *27*(1), 276–284.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37.

Fazel-Zarandi, M., & Fox, M. S. (2012). An ontology for skill and competency management. In *Fois* (pp. 89–102).

Fazio, L. S. (1985). The delphi : Education and assessment in institutional goal setting. *Assessment and Evaluation in Higher Education*, *10*(2), 147–157.

Ferré, S. (2017). Sparklis : An expressive query builder for sparql endpoints with guidance

in natural language. *Semantic Web*, *8*(3), 405–418.

Fischer, G., & Stevens, C. (1990). *Information access in complex, poorly structured information spaces* (Rapport technique). COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE.

Fisher, D. H., Pazzani, M. J., & Langley, P. (2014). *Concept formation : Knowledge and experience in unsupervised learning*. Morgan Kaufmann.

Fornel, M. d. (1990). Qu'est-ce qu'un expert ? connaissances procédurale et déclarative dans l'interaction médicale. *Réseaux. Communication-Technologie-Société*, *8*(43), 59–80.

Fox, M. S., Barbuceanu, M., & Gruninger, M. (1996). An organisation ontology for enterprise modeling : Preliminary concepts for linking structure and behaviour. *Computers in industry*, *29*(1-2), 123–134.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2), 131–163.

Gandon, F., Corby, O., & Faron-Zucker, C. (2012). *Le web sémantique : Comment lier les données et les schémas sur le web ?* Dunod.

Garnine, N. (2020). Système de recommandation basé sur la détection de communautés.

Geer, A. (2021). Learning earth system models from observations : machine learning or data assimilation ? *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200089.

Geller, J. (2009). Ontologies and medical terminologies. In *Encyclopedia of data warehousing and mining, second edition* (pp. 1463–1469). IGI Global.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow : Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gevorkyan, M. N., Demidova, A. V., Demidova, T. S., & Sobolev, A. A. (2019). Review and comparative analysis of machine learning libraries for machine learning. *Discrete and Continuous Models and Applied Computational Science*, *27*(4), 305–315.

Ghazanfar, M. A., & Prügel-Bennett, A. (2014). Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications*, *41*(7), 3261–3275.

Gherabi, C. E. (2018). *Détection des spams se basant sur les techniques de classification*

(Thèse de doctorat non publiée). UNIVERSITE MOHAMED BOUDIAF-M'SILA FACULTE DES MATHEMATIQUES ET DE L ....

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In *Transfer of learning* (pp. 9–46). Elsevier.

Gijana, A. P. (2011). *Assessing challenges in public appointments and recruitment processes in chris hani district municipality : A case study of human resource department in lukhanji local municipality (2008-2010* (Thèse de doctorat non publiée). University of Fort Hare Fort Hare.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–71.

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste : A constant time collaborative filtering algorithm. *information retrieval*, *4*(2), 133–151.

Gómez-Pérez, A. (2004). Ontology evaluation. In *Handbook on ontologies* (pp. 251–273). Springer.

Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system : Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 13.

Gottwald, G. A., & Reich, S. (2021). Supervised learning from noisy observations : Combining machine-learning techniques with data assimilation. *Physica D : Nonlinear Phenomena*, *423*, 132911.

Green, K. C., Armstrong, J. S., & Graefe, A. (2007). Methods to elicit forecasts from groups : Delphi and prediction markets compared.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199–220.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing ? *International journal of human-computer studies*, *43*(5-6), 907–928.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International journal of human-computer studies*, *43*(5-6), 625–640.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. In *Otm confederated international conferences" on the move to meaningful internet systems"* (pp. 986–996).

Gupta, J., & Gadge, J. (2015). Performance analysis of recommendation system based on collaborative filtering and demographics. In *2015 international conference on communication, information & computing technology (iccict)* (pp. 1–6).

Gupta, R., & Singh, M. (2017). Nonclassical symmetries and similarity solutions of variable coefficient coupled kdv system using compatibility method. *Nonlinear Dynamics*, *87*(3), 1543–1552.

Gupta, U. G., & Clarke, R. E. (1996). Theory and applications of the delphi technique : A bibliography (1975–1994). *Technological forecasting and social change*, *53*(2), 185–211.

Halgamuge, S. K., & Wang, L. (2005). *Classification and clustering for knowledge discovery* (Vol. 4). Springer Science & Business Media.

Hamers, L., et al. (1989). Similarity measures in scientometric research : The jaccard index versus salton's cosine formula. *Information Processing and Management*, *25*(3), 315–18.

Hammond, G. S. (1955). A correlation of reaction rates. *Journal of the American Chemical Society*, *77*(2), 334–338.

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets : History and context. *Acm transactions on interactive intelligent systems (tiis)*, *5*(4), 1–19.

Harper, F. M., Li, X., Chen, Y., & Konstan, J. A. (2005). An economic model of user rating in an online recommender system. In *International conference on user modeling* (pp. 307–316).

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., ... others (2018). Applied machine learning at facebook : A datacenter infrastructure perspective. In *2018 ieee international symposium on high performance computer architecture (hpca)* (pp. 620–629).

Henderson, J. C., Thomas, J. B., & Venkatraman, N. (1992). Making sense of it–strategic alignment and organizational context.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5–53.

Hirst, G., St-Onge, D., et al. (1998). Lexical chains as representations of context for the

detection and correction of malapropisms. *WordNet : An electronic lexical database*, *305*, 305–332.

Hitt, M. A., Tyler, B. B., Hardee, C., & Park, D. (1995). Understanding strategic intent in the global marketplace. *Academy of Management Perspectives*, *9*(2), 12–19.

Hunt, J., Myers, J., & Myers, L. (2019). Improving earnings predictions with machine learning. *Unpublished working paper*.

Hwang, M. I., & Lin, J. W. (1999). Information dimension, information overload and decision quality. *Journal of information science*, *25*(3), 213–218.

Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., ... others (2019a). Reinforcement learning for slate-based recommender systems : A tractable decomposition and practical methodology. *arXiv preprint arXiv :1905.12767*.

Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., ... Boutilier, C. (2019b). Slateq : A tractable decomposition for reinforcement learning with recommendation sets.

Ikeda, T., & Stephens, M. (1998). Some characteristics of students'approaches to mathematical modelling in the curriculum based on pure mathematics. *Journal of science education in Japan*, *22*(3), 142–154.

Isaac, H., & Volle, P. (2014). *E-commerce : de la stratégie à la mise en oeuvre opérationnelle*. Pearson Education France.

Jadhav, S. D., & Channe, H. (2016). Efficient recommendation system using decision tree classifier and collaborative filtering. *Int. Res. J. Eng. Technol*, *3*(8), 2113–2118.

Jain, M. (1985). Fifth order implicit multipoint method for solving equations. *BIT Numerical Mathematics*, *25*(1), 250–255.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Jones, M., & Alani, H. (2006). Content-based ontology ranking.

Karatzoglou, A., Baltrunas, L., & Shi, Y. (2013). Learning to rank for recommender systems. In *Proceedings of the 7th acm conference on recommender systems* (pp. 493–494).

Kassab, R. (2009). *Analyse des propriétés stationnaires et des propriétés émergentes dans les flux d'informations changeant au cours du temps* (Thèse de doctorat non

publiée).

Kassel, G. (2018). Ontologie de l'action et formes logiques des phrases d'action : de nouvelles perspectives. *Actes des 12èmes Journées d'Intelligence Artificielle Fondamentale, Amiens*, 13–15.

Kastein, M. R., Jacobs, M., Van Der Hell, R. H., Luttik, K., & Touw-Otten, F. W. (1993). Delphi, the issue of reliability : a qualitative delphi study in primary health care in the netherlands. *Technological forecasting and social change*, *44*(3), 315–323.

Kembellec, G., Chartron, G., & Saleh, I. (2014). *Les moteurs et systèmes de recommandation*. ISTE Group.

Klass, U., Dietsche, W., Von Klitzing, K., & Ploog, K. (1991). Imaging of the dissipation in quantum-hall-effect experiments. *Zeitschrift für Physik B Condensed Matter*, *82*(3), 351–354.

Konstan, J. A., Riedl, J., Borchers, A., & Herlocker, J. L. (1998). Recommender systems : A grouplens perspective. In *Recommender systems : Papers from the 1998 workshop (aaai technical report ws-98-08)* (pp. 60–64).

Koren, Y. (2009). The bellkor solution to the netflix grand prize. *Netflix prize documentation*, *81*(2009), 1–10.

Kumar, S., de A. e Silva, J., Wani, M. Y., Dias, C. M., & Sobral, A. J. (2016). Studies of carbon dioxide capture on porous chitosan derivative. *Journal of Dispersion Science and Technology*, *37*(2), 155–158.

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems–a survey. *Knowledge-Based Systems*, *123*, 154–162.

Lam, S. K., & Riedl, J. (2004). Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on world wide web* (pp. 393–402).

Larreche, J.-C., & Moinpour, R. (1983). Managerial judgment in marketing : The concept of expertise. *Journal of Marketing Research*, *20*(2), 110–121.

Leacock, C., & Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, *49*(2), 265–283.

Lecomte, A., & Quatrini, M. (2010). Pour une étude du langage via l'interaction : dialogues et sémantique en ludique. *Mathématiques et sciences humaines. Mathematics and social sciences*(189), 37–67.

Lecun, Y. (2016). Les enjeux de la recherche en intelligence artificielle. *Interstices*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Leleu-Merviel, S., & Useille, P. (2008). *Quelques révisions du concept d'information.* Lavoisier.

Levesque, H. J. (1986). Knowledge representation and reasoning. *Annual review of computer science*, *1*(1), 255–287.

Levy, M., & Bosteels, K. (2010). Music recommendation and the long tail. In *1st workshop on music recommendation and discovery (womrad), acm recsys, 2010, barcelona, spain.*

Li, H., Wu, D., Tang, W., & Mamoulis, N. (2015). Overlapping community regularization for rating prediction in social recommender systems. In *Proceedings of the 9th acm conference on recommender systems* (pp. 27–34).

Li, Q., & Kim, B. M. (2003). An approach for combining content-based and collaborative filters. In *Proceedings of the sixth international workshop on information retrieval with asian languages-volume 11* (pp. 17–24).

Lian, J., Zhang, F., Hou, M., Wang, H., Xie, X., & Sun, G. (2017). Practical lessons for job recommendations in the cold-start scenario. In *Proceedings of the recommender systems challenge 2017* (pp. 1–6).

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451–461.

Lin, D., et al. (1998). An information-theoretic definition of similarity. In *Icml* (Vol. 98, pp. 296–304).

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations : Item-to-item collaborative filtering. *IEEE Internet computing*(1), 76–80.

Linstone, H. A., Turoff, M., et al. (1975). *The delphi method*. Addison-Wesley Reading, MA.

Liu, C.-H. S., Su, C.-S., Gan, B., & Chou, S.-F. (2014). Effective restaurant rating scale development and a mystery shopper evaluation approach. *International Journal of Hospitality Management*, *43*, 53–64.

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based

on click behavior. In *Proceedings of the 15th international conference on intelligent user interfaces* (pp. 31–40).

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

Lu, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., & Zhou, T. (2012). Recommender systems. *Physics reports*, *519*(1), 1–49.

Luo, X., & Seyedian, M. (2003). Contextual marketing and customer-orientation strategy for e-commerce : an empirical analysis. *International Journal of Electronic Commerce*, *8*(2), 95–118.

Luu, N. D., & Vasavda, D. (2020). Job finder.

Maltz, D., & Ehrlich, K. (1995). Pointing the way : active collaborative filtering. In *Chi* (Vol. 95, pp. 202–209).

Marcus, R. (1988). Semiclassical wave packets in the angle representation and their role in molecular dynamics. *Chemical physics letters*, *152*(1), 8–13.

Marsland, S. (2014). *Machine learning : an algorithmic perspective*. Chapman and Hall/CRC.

Martens, H. H. (1959). Two notes on machine "learning". *Information and Control*, *2*(4), 364–379.

Mbaye, B. (2018). Recommender system : Collaborative filtering of e-learning resources. *International Association for Development of the Information Society*.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., ... Riedl, J. (2002). On the recommending of citations for research papers. In *Proceedings of the 2002 acm conference on computer supported cooperative work* (pp. 116–125).

Middleton, S. E., Alani, H., & De Roure, D. C. (2002). Exploiting synergy between ontologies and recommender systems. *arXiv preprint cs/0204012*.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 54–88.

Miller, G. A. (1998). *Wordnet : An electronic lexical database*. MIT press.

Mohammadian Mahmoudi Tabar, M., Sotoodeh, S., & Boudlaie, H. (2021). Identifying the effective factors of innovative marketing in smes in the it industry. *Journal of Entrepreneurship Development*, *14*(1), 81–98.

Molnar, C. (2019). *Interpretable machine learning*. Lulu. com.

Montaner, M., López, B., & De La Rosa, J. L. (2003). A taxonomy of recommender agents on the internet. *Artificial intelligence review*, *19*(4), 285–330.

Moreno, A., Valls, A., Isern, D., Marin, L., & Borràs, J. (2013). Sigtur/e-destination : ontology-based personalized recommendation of tourism and leisure activities. *Engineering applications of artificial intelligence*, *26*(1), 633–651.

Morin, E. (1992). *La méthode. 3 : La connaissance de la connaissance : Anthropologie de la connaissance*. Editions du Seuil.

Morin, E. (2008). *L'esprit du temps*. Armand Colin.

Myllynen, S., Suominen, I., Raunio, T., Karell, R., & Lahtinen, J. (2021). Developing and implementing ai-based classifier for requirements engineering. *Journal of Nuclear Engineering and Radiation Science*.

Navarro, M. O., Piva, A. C., Simionato, A. S., Spago, F. R., Modolon, F., Emiliano, J., ... Andrade, G. (2019). Bioactive compounds produced by biocontrol agents driving plant health. In *Microbiome in plant health and disease* (pp. 337–374). Springer.

Newell, A. (1982). The knowledge level. *Artificial intelligence*, *18*(1), 87–127.

Ngom, A. N. (2015). Étude des mesures de similarité sémantique basées sur les arcs. In *Coria* (pp. 535–544).

Ngom, M. A. N. (2018). *Docteur en informatique* (Thèse de doctorat non publiée). Université de Bordeaux.

Nguyen, A.-T., Denos, N., & Berrut, C. (2006). Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride..

Niang, C. A. T. (2013). *Vers plus d'automatisation dans la construction de systèmes mediateurs pour le web semantique : une application des logiques de description* (Thèse de doctorat non publiée). Tours.

Ogden, C. K., & Richards, I. A. (1923). The meaning of meaning : A study of the influence of thought and of the science of symbolism.

O'Mahony, M. P., & Smyth, B. (2007). A recommender system for on-line course en-

rolment : an initial study.

Otlet, P. (1934). Traité de documetation : Le livre sur le livre. *Théorie et Pratique, Bruxelles : Editions*.

Ouellet, J., & Tessier, J.-C. (1987). Intelligence artificielle et systèmes experts : Principes et méthodes.

Oufaida, H., & Nouali, O. (2008). Le filtrage collaboratif et le web 2.0. *Document numérique*, *11*(1), 13–35.

Pavlenko, T., & Von Rosen, D. (2001). Effect of dimensionality on discrimination. *Statistics*, *35*(3), 191–213.

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, *13*(5-6), 393–408.

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

Penalva, J., & Montmain, J. (2002). Travail collectif et intelligence collective : les référentiels de connaissances. In *Ipmu'2002, 9th international conference on information processing and management of uncertainty in knowledge-based systems, annecy, france.*

Perugini, S., Gonçalves, M. A., & Fox, E. A. (2004). Recommender systems research : A connection-centric survey. *Journal of Intelligent Information Systems*, *23*(2), 107–143.

Pessiot, J.-F., Truong, T.-V., Usunier, N., Amini, M.-R., & Gallinari, P. (2006). Factorisation en matrices non négatives pour le filtrage collaboratif. In *Coria* (pp. 315–326).

Piamrat, K., Viho, C., Bonnin, J.-M., & Ksentini, A. (2009). Quality of experience measurements for video streaming over wireless networks. In *2009 sixth international conference on information technology : New generations* (pp. 1184–1189).

Piotte, M., & Chabbert, M. (2009). The pragmatic theory solution to the netflix grand prize. *Netflix prize documentation*.

Popescul, A., Pennock, D. M., & Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*

(pp. 437–444).

Quillian, M. R. (1967). Word concepts : A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

Rabinowitz, M., & Glaser, R. (1985). *Cognitive structure and process in highly competent performance.* American Psychological Association.

Rabut, T. (2020). Apprentissage automatique appliqué au bitcoin.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, *19*(1), 17–30.

Rakoto, P. (2005). Caractéristiques de l'information, surcharge d'information et qualité de la prédiction. *Comptabilité-Contrôle-Audit*, *11*(1), 23–38.

Ramkumar, P. N., Haeberle, H. S., Navarro, S. M., Sultan, A. A., Mont, M. A., Ricchetti, E. T., ... Iannotti, J. P. (2018). Mobile technology and telemedicine for shoulder range of motion : validation of a motion-based machine-learning software development kit. *Journal of shoulder and elbow surgery*, *27*(7), 1198–1204.

Rao, K. N. (2008). Application domain and functional classification of recommender systems—a survey. *DESIDOC Journal of Library & Information Technology*, *28*(3), 17–35.

Rastier, F. (1995). La sémantique des thèmes-ou le voyage sentimental. *L'analyse thématique des données textuelles. L'exemple des sentiments, Paris : Didier*, 223–249.

Razmerita, L., Angehrn, A., & Maedche, A. (2003). Ontology-based user modeling for knowledge management systems. In *International conference on user modeling* (pp. 213–217).

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

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–59.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you ?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Roqueplo, P. (1997). *Entre savoir et décision, l'expertise scientifique*. Éditions Quae.

Rowley, J. (2007). The wisdom hierarchy : representations of the dikw hierarchy. *Journal of information science*, *33*(2), 163–180.

Sackman, H. (1974). *Delphi assessment : Expert opinion, forecasting, and group process* (Rapport technique). Rand Corp Santa Monica CA.

Safoury, L., & Salah, A. (2013). Exploiting user demographic attributes for solving cold-start problem in recommender system. *Lecture Notes on Software Engineering*, *1*(3), 303–307.

Sahraoui, K. (2017). *Modélisation sémantique des systèmes de filtrage collaboratif hybride (sfch) à base ontologique* (Thèse de doctorat non publiée). ESI.

Salton, G. (1989). Automatic text processing : The transformation, analysis, and retrieval of. *Reading : Addison-Wesley*, *169*.

Sarwar, B. M., Karypis, G., Konstan, J. A., Riedl, J., et al. (2001). Item-based collaborative filtering recommendation algorithms. *Www*, *1*, 285–295.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291–324). Springer.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international acm sigir conference on research and development in information retrieval* (pp. 253–260).

Schilit, B. N., & Theimer, M. M. (1994). Disseminating active mop infonncition to mobile hosts. *IEEE network*.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Ecai* (Vol. 16, p. 1089).

Shardanand, U., & Maes, P. (1995). Social information filtering : algorithms for automating" word of mouth". In *Chi* (Vol. 95, pp. 210–217).

Sheth, B., & Maes, P. (1993). Evolving agents for personalized information filtering. In *Proceedings of 9th ieee conference on artificial intelligence for applications* (pp.

345–352).

Sicilia, M.-Á., Rodríguez, D., García-Barriocanal, E., & Sánchez-Alonso, S. (2012). Empirical findings on ontology metrics. *Expert Systems with Applications*, *39*(8), 6706–6711.

Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv :1310.8059*.

Slovic, P. (1972). Psychological study of human judgment : Implications for investment decision making. *The Journal of Finance*, *27*(4), 779–799.

Slovic, P., & Lichtenstein, S. (1971). Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational behavior and human performance*, *6*(6), 649–744.

Soboroff, I., & Nicholas, C. (1999). Combining content and collaboration in text filtering. In *Proceedings of the ijcai* (Vol. 99, pp. 86–91).

Sohail, S. S., Siddiqui, J., & Ali, R. (2014). User feedback scoring and evaluation of a product recommendation system. In *2014 seventh international conference on contemporary computing (ic3)* (pp. 525–530).

Soualah-Alila, F., Nicolle, C., & Mendes, F. (2014). Une approche web sémantique et combinatoire pour un systéme de recommandation sensible au contexte appliqué à l'apprentissage mobile. In *11 ème édition de l'atelier fouille de données complexes.*

Souilah, S. (2019). Nouvelle méthode de démarrage à froid pour les systèmes de recommandation.

Sowa, J. F. (2000). Ontology, metadata, and semiotics. In *International conference on conceptual structures* (pp. 55–81).

Steck, H. (2013). Evaluation of recommendations : rating-prediction and ranking. In *Proceedings of the 7th acm conference on recommender systems* (pp. 213–220).

Steels, L. (2002). A bibliography of publications of luc steels.

Stefani, A., & Strapparava, C. (1999). Exploiting nlp techniques to build user model for web sites : the use of wordnet in siteif project. In *Proc. 2nd workshop on adaptive systems and user modeling on the www.*

Stenmark, D. (2002). Information vs. knowledge : The role of intranets in knowledge management. In *Proceedings of the 35th annual hawaii international conference on*

system sciences (pp. 928–937).

Stolze, M., & Rjaibi, W. (2001). Towards scalable scoring for preference-based item recommendation. *IEEE Data Eng. Bull.*, *24*(3), 42–49.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international conference on information and knowledge management* (pp. 67–74).

Syed, Z., Finin, T., Mulwad, V., Joshi, A., et al. (2010). Exploiting a web of semantic data for interpreting tables. In *Proceedings of the second web science conference.*

Tadlaoui, M., George, S., & Sehaba, K. (2015). Approche pour la recommandation de ressources pédagogiques basée sur les liens sociaux. In *7ème conférence sur les environnements informatiques pour l'apprentissage humain (eiah 2015)* (pp. 192–203).

Tahiraly, R. T. (2014). *Conception et évaluation d'un système décisionnel informatisé basé sur le raisonnement des experts élaborant les guides de bonnes pratiques en antibiothérapie empirique* (Thèse de doctorat non publiée). Université Paris-Nord-Paris XIII.

Ticha, S. B. (2015). *Recommandation personnalisée hybride* (Thèse de doctorat non publiée). Université de Lorraine.

Tran, T., & Cohen, R. (2000). Hybrid recommender systems for electronic commerce. In *Proc. knowledge-based electronic markets, papers from the aaai workshop, technical report ws-00-04, aaai press* (Vol. 40).

Tsang, M., Cheng, D., Liu, H., Feng, X., Zhou, E., & Liu, Y. (2020). Feature interaction interpretability : A case for explaining ad-recommendation systems via neural interaction detection. *arXiv preprint arXiv :2006.10966*.

Tsang, M., Cheng, D., & Liu, Y. (2017). Detecting statistical interactions from neural network weights. *arXiv preprint arXiv :1705.04977*.

Tsuchiya, S. (1995). Commensurability, a key concept of business reengineering. In *Proceedings 3rd international symposium on the management of industrial and corporate knowledge ismick* (Vol. 95, pp. 81–87).

Vellinga, N. E. (2017). From the testing to the deployment of self-driving cars : Legal challenges to policymakers on the road ahead. *Computer Law & Security Review*,

33(6), 847–863.

Vernette, E. (1994). La méthode delphi : une aide à la prévision marketing. *Décisions Marketing*, 97–101.

Vernette, E. (1997). *Evaluation de la validité prédictive de la méthode delphi-leader, 13ème congrés international de l'association française de marketing, 13.* Toulouse.

Vernette, E., & Marketing, C. (2007). Une nouvelle méthode de groupe pour interpréter le sens d'une expérience de consommation :≪l'album on-line≫(aol). *Actes des 12èmes Journées de Recherche en Marketing de Bourgogne*, *19*.

Vieira, L. (2014). Les réseaux et l'humain. exploration de la genèse d'une nouvelle expertise. *Sciences de la société*(91), 12–25.

Viola, R., Emonet, R., Habard, A., Metzler, G., Riou, S., & Sebban, M. (2019). Une version corrigée de l'algorithme des plus proches voisins pour l'optimisation de la f-mesure dans un contexte déséquilibré. In *Conférence sur l'apprentissage automatique (cap 2019).*

Vogel, C. (1988). *Génie cognitif.* Masson.

Vogt, C. C., Cottrell, G. W., Belew, R. K., & Bartell, B. T. (1997). Using relevance to train a linear mixture of experts. *NIST SPECIAL PUBLICATION SP*, 503–516.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems.

Webster, J., & Trevino, L. K. (1995). Rational and social theories as complementary explanations of communication media choices : Two policy-capturing studies. *Academy of Management journal*, *38*(6), 1544–1572.

Welty, G. (1971). A critique of the delphi technique. *Proceedings of the American Statistical Association. Washington DC*, 377–382.

Westenberg, M. R., & Koele, P. (1994). Multi-attribute evaluation processes : Methodological and conceptual issues. *Acta Psychologica*, *87*(2-3), 65–84.

Wielinga, B. J., Schreiber, A. T., & Breuker, J. A. (1992). Kads : A modelling approach to knowledge engineering. *Knowledge acquisition*, *4*(1), 5–53.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on association for computational linguistics* (pp. 133–138).

Xu, Z. (2007). Some similarity measures of intuitionistic fuzzy sets and their applications to multiple attribute decision making. *Fuzzy Optimization and Decision Making*,

6(2), 109–121.

Xuan, Q., Ruan, Z., & Min, Y. (2021). *Graph data mining : Algorithm, security and application.* Springer.

Ye, B. K., Tu, Y. J. T., & Liang, T. P. (2019). A hybrid system for personalized content recommendation. *Journal of Electronic Commerce Research*, *20*(2), 91–104.

Zargayouna, H. (2005). *Indexation sémantique de documents xml* (Thèse de doctorat non publiée). Paris 11.

Zedeck, S. (1977). An information processing model and approach to the study of motivation. *Organizational Behavior and Human Performance*, *18*(1), 47–77.

Zeleny, M. (1987). Management support systems : towards integrated knowledge management. *Human systems management*, *7*(1), 59–70.

Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, *7*(1), 439–457.

Zoungrana, W.-B. A. B. (2020). *Application des algorithmes d'apprentissage automatique pour la détection de défauts de roulements sur les machines tournantes dans le cadre de l'industrie 4.0* (Thèse de doctorat non publiée). Université du Québec à Chicoutimi.

Table des figures

Liste des tableaux

Abréviations

Liste des abréviations

SRSE système de recommandation sémantique enrichi

SR système de recommandation

MoPRD module de préparation et de représentation des données

MoP module de prédiction

MoCI module de classification des items

CoSD composant de sélection des données

CoECD composant d'extraction des connaissances à partir des données

CoRCHD composant de représentation des connaissances et d'homogénéisation des données

CoFCPV composant de formation de communautés de points de vente

ECD extraction des connaissances à partir des données

FD Fouille de données

Annexes

Annexe A

Algorithm 3 Importations des data-sets

• Importation des bibliothèques nécessaires

import seaborn as sb

import pandas as pd

import tensorflow as tf

from tensorflow import keras

from sklearn.neighbors

import KNeighborsClassifier

• Chargement des data-sets dans les fichiers CSV

columns_names=['id','*company_name*','skills','skills-stories']

• Chargement des données d'entrainement

training_matrice_path =$tf.keras.utils.get\_file$("matrices.csv")

• Chargement des données de test

test_matrice_path= $tf.keras.utils.get\_file$("matrices.csv")

Algorithm 4 Lecture et concaténation des données d'entrainement

• Lecture des données d'entrainement

matriceTraining = *pd.read_csv*(*training_matrice_path*, names=*columns_names*, header=0)

• Lecture des données d'entrainement

matriceTest = *pd.read_csv*(*test_matrice_path*, names=*columns_names*, header=0)

matriceTest = matriceTest[matriceTest[*test_matrice_path*]]

• Concaténation des données d'entrainement et celles des tests

matrices_dataset = pd.concat([matriceTraining, matriceTest], axis=0)

matrices_dataset.head()

matrices_dataset.describe()

Algorithm 5 Classification des items

• Importation des bibliothèques nécessaires

import seaborn as sb

import pandas as pd

import tensorflow as tf

from tensorflow import keras

from tensorflow.estimator

import LinearClassifier

• Chargement des data-sets dans les fichiers CSV

$columns\_names = ['id','company\_name','skills','skills - stories']$

• Chargement des données d'entrainement

$training\_data\_path{=}tf.keras.utils.get\_file("skills\_training.csv")$

• Chargement des données de test

$test\_data\_path = tf.keras.utils.get\_file("skeel\_test.csv")$

• Lecture des données d'entrainement

$training = pd.read\_csv(training\_data\_path, names = columns\_names, header = 0)$

• Lecture des données d'entrainement

$test = pd.read\_csv(test\_data\_path, names = columns_names, header = 0)$

$test = test[test[test\_data\_path]]$

• Concaténation des données d'entrainement et celles des tests

$skill\_dataset = pd.concat([training, test], axis = 0)$

$skill\_dataset.describe()$

• Corrélation entre les données

$correlation\_data = skill\_dataset.corr()$

$correlation\_data.style.background\_gradient(cmap = 'coolwarm', axis = None)$

Algorithm 6 Entrainement du modèle

• Statistiques

$stats = skill\_dataset.describe()$

$skill\_stats = stats.transpose()$

$skill\_stats$

$X\_data = skill\_dataset[[kforkinskill\_dataset.columnsifknotin['Skills']]]$

$Y\_data = skill\_dataset[['Skills']]$

• Entrainement et normalisation

$training\_features, test\_features = train\_test_split(X\_data, Y_data, test\_size = 0.7)$

$defnormalize(x) : normed\_train\_features = normalize(training\_features)$

$normed\_test\_features = normalize(test\_features)definput\_feed\_function() :$

$dataset = tf.data.Dataset.from\_tensor\_slices()$

$ifshuffle : dataset = dataset.shuffle(2000)$

$dataset = dataset.batch(32).repeat(num\_of\_epochs)$

$returndataset$

$returninput\_feed\_function$

$train\_feed\_input = feed\_input(normed\_train\_features)$

$train\_feed\_input\_testing = feed\_input(normed\_train\_features, 1, false)$

$test_feed_input = feed_input(normed\_test\_features, test\_labels, 1, false)$

• Entrainement du modèle

$feature\_columns\_numeric =$

$[tf.feature\_column.numeric\_column(k)forkintraining\_features.columns]$

$logistic\_model = LinearClassifier(feature\_columns = feature\_columns\_numeric)$

$logistic\_model.train(train\_feed\_input)$

Algorithm 7 Prédiction

$train\_predictions = logistic\_model.predict(train\_feed\_input\_testing)$

$test\_predictions = logistic\_model.predict(test\_feed\_input)$

$train\_predictions\_series =$

$pd.Series([p['classes'][0].decode("utf-8")forpintrain\_predictions])$

$test\_predictions\_series =$

$pd.Series([k['classes'][0].decode("utf-8")forkintest\_predictions])$

$train\_predictions\_df =$

$pd.DataFrame(train\_predictions\_series, columns = ['predictions'])$

$test\_predictions\_df =$

$pd.DataFrame(test\_predictions\_series, columns = ['predictions'])$

$train\_predictions\_df.reset\_index(drop = True, inplace = True)$

$test\_predictions\_df.reset\_index(drop = True, inplace = True)$

Algorithm 8 Classification finale

$defcalculate\_binary\_class\_scores(y_truey_pred) :$

$relevant = relevant\_score(y\_true, y\_pred.astype('int64'))$

$precision = precision\_score(y\_true, y\_pred.astype('int64'))$

$returnrelevant, precision$

$train\_relevant\_score, train\_precision\_score$

$= calculate\_binary\_class\_scores(training\_labels, train\_predictions\_series)$

$test\_accuracy\_score, test\_precision\_score$

$= calculate\_binary\_class\_scores(', test\_predictions\_series)$

Algorithm 9 Notre algorithme de « Page Rank »

import numpy as np

import pandas as pd

import Tensorflow as tf

import scipy as sc

from fractions import Fraction

def $display_format(my_vector, my_decimal)$:

$returnnp.round((my_vector).astype(np.float), decimals = my_decimal)$

$data = Fraction(tf.keras.utils.get_file("matrices.csv"))$

$Mat = np.matrix(tf.keras.utils.get_file("matrices.csv"))$

$-----$

$-----$

$-----$

$Ex[:] = data$

$beta = 0.7$

$K = beta * Mat + ((1 - beta) * Ex)$

$-----$

$-----$

$-----$

$trans = np.matrix(tf.keras.utils.get_file("matrices.csv"))$

$trans = np.transpose(trans)$

$previous_trans = trans$

$foriiinrange(1, 500)$:

$trans = K * trans$

$if(previous_r == r).all()$:

$break$

$previousTrans = trans$

$-----$

$-----$

Annexe B

La liste des requêtes pour rendre anonyme les données des différents points de vente :

— **anonymisation du nom des points de vente**

UPDATE target SET target.name = 'pointDeVente'+target.id ;

— **anonymisation du nom des zones des points de vente**

UPDATE level SET level.name = 'secteur'+level.id ;

— **anonymisation du nom réseau des points de vente**

UPDATE network SET network.name = 'reseau'+network.id ;