MÉMOIRE

Présenté par : JACQUET Noémie

Dans le cadre du **Certificat de Spécialité :**
IODAA AgroParisTech

Stage effectué du : 06/03/2023 au 21/09/2023
À : AgroParisTech, Département MMIP, UMR MIA-Paris-Saclay,
22 place de l'agronomie, 91123 Palaiseau

Sur le **thème** :
EXERSYS, un système de recommandation appliqué à la nutrition, combinant graphes de connaissances et apprentissage automatique

Pour l'obtention du :
CERTIFICAT DE SPECIALITE

Enseignant/e-référent : Madame Liliana IBANESCU

Maître de stage : Madame Cristina MANFREDOTTI

Soutenu le : 05/09/2023

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je tiens à remercier chaleureusement ma Maître de stage Madame Cristina Manfredotti pour son encadrement, ses précieux conseils, sa grande disponibilité et son écoute. Je la remercie aussi d'avoir partagé régulièrement avec moi sa vision inspirante du sujet EXERSYS, sujet dont elle est à l'origine.

Je tiens à remercier vivement Monsieur Vincent Guigue pour ses recommandations, ses explications sur les méthodes existantes et son aide précieuse sur la programmation les nombreuses fois où je l'ai sollicité, sa grande disponibilité.

Madame Cristina Manfredotti et Monsieur Vincent Guigue m'ont aidée à avoir une vision globale sur le sujet et ses enjeux en parallèle des aspects techniques. Ils m'ont guidée sur la démarche scientifique, aidée sur l'implémentation de nouvelles méthodes et l'interprétation des résultats. J'ai énormément (le mot n'est pas assez fort) appris auprès d'eux. Je les en remercie encore.

Je remercie également les acteurs du projet EXERSYS : Madame Fatiha Saïs, Monsieur Paolo Viappiani, Monsieur Stéphane Dervaux pour leurs conseils, leurs retours inspirants et décisifs pour le projet à chaque réunion de travail.

Je remercie Monsieur Ayoub Hammal, stagiaire pour ses riches travaux qui sont cités dans mon rapport et qui m'ont permis de tester les recommandations émises par mon modèle. Je le remercie aussi pour les programmes qu'il a écrits pour connecter nos deux « systèmes » et pour son esprit d'équipe.

Je remercie mon Enseignant référent Madame Liliana Ibanescu, pour ses conseils avisés, la relecture de mon rapport et sa disponibilité.

Je remercie mes Professeurs Monsieur Antoine Cornuéjols et Madame Christine Martin de m'avoir acceptée en formation IODAA.

Je remercie enfin mon collègue de bureau Amine pour les moments partagés.

Résumé

Favoriser de bonnes habitudes alimentaires est un enjeu majeur de santé publique. Un système de recommandation qui tient compte à la fois des préférences des utilisateurs et des contraintes nutritionnelles peut jouer un rôle déterminant pour accroître l'adhésion aux recommandations nutritionnelles Le stage dans le cadre du projet EXERSYS a pour objectif de concevoir un tel système. L'objectif est de développer un système de recommandation personnalisée basé sur l'apprentissage des préférences des utilisateurs (axe 1 du projet) et d'intégrer à la recommandation des contraintes nutritionnelles en exploitant les graphes de connaissances (axe 2 du projet).

Ce projet se distingue par son caractère novateur, car la plupart des systèmes de recommandation alimentaire se concentrent sur un seul aspect et recommandent des aliments individuels (généralement par filtrage collaboratif), plutôt que des repas complets. Une autre force de ce projet réside dans sa capacité à intégrer des informations contextuelles spécifiques à chaque utilisateur, comme les allergies ou le végétarisme.

Pour l'axe 1 qui se concentre sur les recommandations par apprentissage automatique, les préférences alimentaires sont issues des données de consommation de l'enquête INCA2. Le premier modèle exploré appelé FiltreCollab repose sur une approche de type filtrage collaboratif. Il se base sur l'apprentissage d'un espace de représentation des aliments où des aliments proches sont substituables (entrainement de l'algorithme Word2Vec sur les repas INCA2), le regroupement en catégories d'aliments substituables, la modélisation de l'utilisateur au sein des catégories et le tirage d'aliments par catégorie selon une loi multinomiale adaptée qui tient compte des distances apprises. La cohérence de la recommandation de repas est assurée par l'application de règles d'association et d'exclusion entre aliments.
Ce modèle FiltreCollab a démontré sa performance pour le petit déjeuner, notamment en termes de log-vraisemblance et de rang moyen réciproque des aliments consommés lors d'une journée masquée, surpassant ainsi un modèle de recommandation aléatoire par catégorie. Cependant, l'application du modèle au déjeuner/diner a révélé certaines limites, notamment l'échec du clustering en catégories d'aliments substituables, et une performance moindre (log-vraisemblance équivalente à un modèle aléatoire par catégorie, bien que le rang moyen réciproque demeure supérieur).
Ces résultats suggèrent un apprentissage partiel où un ordonnancement des aliments a été appris mais pas des distances absolues. Nous faisons l'hypothèse que cela est dû à une plus grande variété d'aliments consommés (proportion d'aliments nouveaux) et à un nombre accru de combinaisons d'aliments pour le déjeuner/diner, ainsi qu'à l'existence d'une séquentialité dans le repas (entrée, plats, dessert) qui n'est pas prise en compte par ce modèle.

Toujours dans le cadre de l'axe 1, en utilisant l'ordre de consommation des aliments disponible dans l'enquête INCA2, nous cherchons à exploiter le caractère séquentiel du déjeuner/diner dans un second modèle. Ce modèle appelé GenSeqRNN est un réseau de neurones récurrent entrainé sur les repas INCA2, avec l'objectif de maximiser la probabilité du prochain aliment ; il est ensuite utilisé pour générer des séquences de repas.
Le modèle GenSeqRNN a une performance encourageante en termes de taux de bonne prédiction du prochain aliment (« accuracy » dans le top 3). De plus il semble améliorer la cohérence des repas générés, une cohérence difficile à obtenir dans le premier modèle en raison du grand nombre de règles d'association/exclusion. Contrairement au modèle FiltreCollab, le modèle GenSeqRNN n'est pas personnalisé mais des pistes de personnalisation ont été identifiées. Bien qu'il semble moins adapté au petit déjeuner qui ne

comporte pas de moments de consommation, il mérite d'être testé sur ce repas également.

Dans le cadre de l'axe 2 (intégration des contraintes nutritionnelles), une base de connaissance a été créée (travaux d'un autre stagiaire) pour modéliser les connaissances sur les aliments et utilisateurs et pour stocker les règles, et un processus de vérification des règles a été défini. La connexion de l'axe 2 à l'axe 1 a été examinée en appliquant des contraintes nutritionnelles aux recommandations de petit déjeuner générées par le modèle FiltreCollab (filtrage collaboratif Word2Vec).
Les tests préliminaires montrent qu'il est possible d'émettre des recommandations valides respectant les contraintes définies pour un utilisateur telles que les allergies, le végétarisme, ou encore les règles d'association/exclusion des aliments. L'ambition est d'étendre cette méthode pour intégrer d'autres éléments de contexte à l'avenir.

En conclusion les différents modèles testés répondent individuellement à différents enjeux du projet : le modèle FiltreCollab se concentre sur la personnalisation de la recommandation, un élément clé de satisfaction utilisateur, tandis que le modèle GenSeqRNN exploite la séquentialité pour améliorer la performance des recommandations de déjeuner/diner. La liaison avec l'axe 2 illustre la possibilité de prendre en compte les contraintes spécifiques des utilisateurs, un enjeu clé en nutrition.
Le défi majeur à relever pour la suite du projet consiste à combiner et adapter les différentes méthodes afin de parvenir à un système de recommandation "global" de repas à l'échelle de la journée et au-delà.

Résumé

Promoting healthy eating habits is a significant public health concern. A recommendation system that takes into account both user preferences and nutritional constraints can play a crucial role in enhancing adherence to nutritional guidelines. My internship and EXERSYS project aim to design such a system. The objective is to develop a personalized recommendation system based on user preference learning (axis 1 of the project) and to integrate nutritional constraints into recommendations through the use of knowledge graphs (axis 2 of the project).

This project is innovative as most food recommendation systems focus on a singular aspect and recommend individual foods (usually through collaborative filtering) rather than complete meals. Another strength of this project lies in its ability to incorporate user-specific contextual information, such as allergies or vegetarian preferences.

For axis 1, which focuses on recommendations through user preference learning, food preferences are derived from consumption data from the INCA2 survey. The first explored model named "FiltreCollab" employs a collaborative filtering approach. It consists of learning a representation space of foods where similar foods are substitutable (training the W2V algorithm on INCA2 meals), clustering into categories of substitutable foods, modeling user preferences within these categories, and selecting foods from each category based on an adapted multinomial law that considers learned distances. Meal recommendation coherence is ensured through the application of association and exclusion rules between foods.
The FiltreCollab model demonstrated its effectiveness for breakfast, particularly in terms of log-likelihood and mean reciprocal rank (MRR) of foods consumed during a hidden day, outperforming a recommendation model with randomly chosen items per category. However, applying the model to lunch/dinner revealed limitations, including the failure of clustering substitutable foods and lower performance (log-likelihood equivalent to model with randomly chosen items per category, although mean reciprocal rank remains superior). These findings suggest partial learning, where a food ordering was learned but not absolute distances. We hypothesize that this is due to a greater variety of consumed foods (proportion of new foods) and an increased number of food combinations for lunch/dinner, as well as the presence of sequence patterns during meal (appetizer, main course, dessert) which are not considered in this model.

Still within the scope of axis 1, using the available food consumption order data from the INCA2 survey, we aim to leverage the sequential nature of lunch/dinner with a second model. This model named "GenSeqRNN" consists of a recurrent neural network trained on INCA2 meals, with the goal of maximizing the probability of the next food item, and it is used to generate meal sequences.
The GenSeqRNN model displays promising performance in terms of accurately predicting the next food item (top 3 accuracy). Additionally, it seems to improve the coherence of generated meals, coherence that was challenging to achieve in the first model due to the numerous association/exclusion rules. Unlike the FiltreCollab model, the GenSeqRNN model is not personalized but routes for personalization have been identified. While it seems less suited to breakfast due to the absence of distinct consumption moments, it is worthwhile to test it for breakfast as well.

Regarding axis 2 (integration of nutritional constraints), a knowledge base has been created (the work of another intern) to model information about foods and users, store rules, and the rule verification process has been established. Connecting axis 2 to axis 1

was examined by applying nutritional constraints to breakfast recommendations generated by FiltreCollab model (collaborative filtering Word2Vec).
Preliminary tests indicate the feasibility of issuing valid recommendations that respect users' constraints such as allergies, vegetarianism, or association/exclusion rules for foods. The ambition is to expand this method to integrate additional elements of context in the future.

In conclusion, the different tested models individually address various project challenges : Filtrecollab model focuses on recommendation personalization, a key element for user satisfaction, while GenSeqRNN model exploits sequence patterns to enhance lunch/dinner recommendation performance. The connection with axis 2 illustrates the potential to consider specific user constraints, a significant aspect in the field of nutrition.
The major challenge ahead is to combine and adapt the various methods to achieve a "global" meal recommendation system that spans the entire day and beyond.

1 Encadrement et environnement du stage

Mon stage se déroule à l'UMR MIA Paris-Saclay (Mathématique et Informatique Appliquées) qui est associée aux tutelles AgroParisTech, INRAE et Université Paris Saclay et regroupe des statisticiens et des informaticiens spécialisés dans la modélisation et l'apprentissage statistique et informatique pour la biologie, l'écologie, l'environnement, l'agronomie et l'agro-alimentaire, et plus précisément au sein du Département MMIP (Modélisation Mathématique Informatique et Physique).

Mon stage s'inscrit dans le projet EXERSYS. Ce projet vise à construire un système de recommandations nutritionnelles, intégrant des données hétérogènes qui vont des données de consommations et préférences alimentaires aux contraintes dictées par les experts, en combinant des méthodes du domaine de l'apprentissage automatique et des graphes de connaissances. Ce projet est supervisé par plusieurs experts étant donné son interdisciplinarité :

- Madame Cristina Manfredotti, Maître de conférences - Ekinocs - AgroParisTech - INRAE - Université Paris-Saclay.
- Monsieur Vincent Guigue, HDR, Professeur - Ekinocs - AgroParisTech - INRAE - Université Paris-Saclay.
- Monsieur Stéphane Dervaux, Ingénieur en informatique - Ekinocs - AgroParisTech - INRAE - Université Paris-Saclay
- Madame Fatiha Saïs, HDR, Professeur - LaHDAK - LISN - Université Paris-Saclay.
- Monsieur Paolo Viappiani, Chercheur – CNRS – LAMSADE – Université Paris-Dauphine.
- Monsieur Nicolas Darcel, HDR, Maître de conférences - PNCA - AgroParisTech - INRAE - Université Paris-Saclay

Compte tenu de l'interdisciplinarité du projet, un encadrement équilibré en Représentation des Connaissances et Raisonnement (Mme Saïs) et Apprentissage automatique (M. Guigue) est crucial. Mme Manfredotti a l'expertise pour faire le pont entre les deux disciplines et en méthodes de calcul en systèmes de recommandation nutritionnelle avec la participation à différents projets liés au domaine de la nutrition (projet ANR SHIFT et une soumission européenne plus d'autres collaborations impliquant le consortium du projet SHIFT). L'expertise de M. Viappiani réside dans la théorie de la décision et la modélisation de l'utilisateur, les algorithmes pour obtenir les préférences des utilisateurs et les techniques de recommandations d'objets structurés. Le soutien de M. Darcel et M. Dervaux est également fondamental pour la pleine réalisation du projet : l'expertise de M. Darcel dans l'application de la nutrition sera utilisée pour construire une base fiable de contraintes nutritionnelles et prototypes utilisateurs du service tandis que l'expertise de M. Dervaux en intégration de données et en ingénierie des ontologies sera d'une grande aide pour la modélisation des connaissances.

Dans mon travail au quotidien, j'ai sollicité régulièrement les conseils et expertises de Madame Manfredotti qui est mon Maître de stage et de Monsieur Guigue.
Des réunions de projet ont eu lieu avec l'ensemble des encadrants à une fréquence bimensuelle.
Un stagiaire étudiant en Master d'Intelligence Artificielle, M. Ayoub Hammal, a également travaillé sur le projet, sur la partie graphes de connaissances, pendant une durée de 2 mois.

2 Objectif et intérêt du travail

2.1 Objectif du stage

Le stage a pour objectif de développer un système de recommandation de repas combinant les techniques de l'apprentissage automatique et des graphes de connaissance. Il s'agit de créer un système de recommandation personnalisée basé sur les préférences des utilisateurs par apprentissage automatique à partir des données de consommations réelles d'individus d'une enquête alimentaire, puis d'intégrer à la recommandation des contraintes nutritionnelles et des règles basées sur des connaissances "expert", par le biais des graphes de connaissances. Le projet a donc été divisé en ces deux axes :

— Axe 1 : élaboration d'un système de recommandation basé sur l'apprentissage automatique et les préférences des utilisateurs

— Axe 2 : intégration à la recommandation des contraintes nutritionnelles avec les graphes de connaissance.

Mon stage de 6 mois porte essentiellement sur l'axe 1.
Le stage de M. Ayoub Hammal a porté sur l'axe 2 avec la création d'une base de connaissances et des processus de vérifications des règles nutritionnelles (Fig. 1).
Le présent rapport présentera donc l'axe 1 (sections 3 et 4) et sa connexion avec l'axe 2 (section 5).

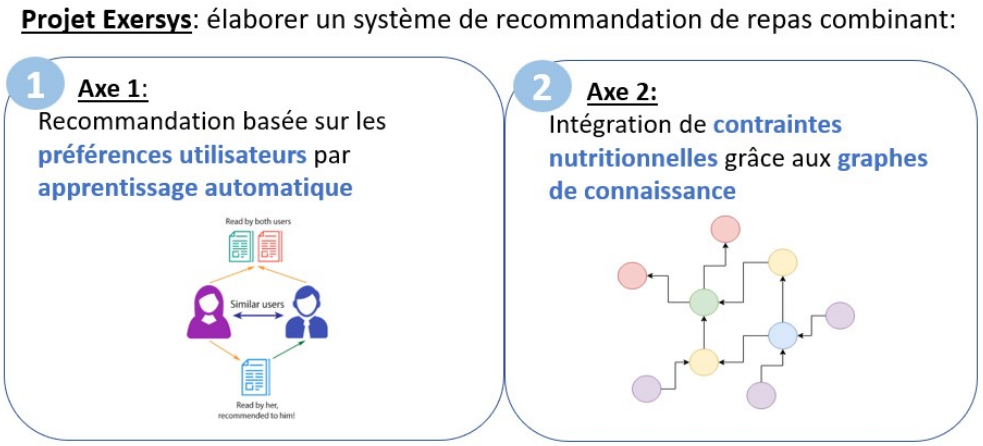

FIGURE 1 – *Objectif du projet EXERSYS*

2.2 Source des données

2.2.1 Données relatives aux préférences des utilisateurs

Concernant les préférences des utilisateurs, les données de consommations alimentaires utilisées sont issues de l'enquête INCA2 [1]. Il s'agit d'une enquête nationale menée de 2006 à 2007 sur la consommation alimentaire individuelle des Français, avec des journaux alimentaires de 7 jours de recueil des consommations pour 4079 individus dont 2624 adultes sur plusieurs mois, en tenant compte de la saisonnalité possible des habitudes alimentaires. Nous avons exclu les consommations des enfants et adolescents de moins de 18 ans.
On considère seulement le fait que les aliments aient été consommés ou non au cours d'un repas,

1. https ://www.data.gouv.fr/fr/datasets/donnees-de-consommations-et-habitudes-alimentaires-de-letude-inca-2-3/

et non les quantités d'aliments qui ont été consommées. Ces données présentent l'avantage d'être structurées, de bonne qualité, et de comporter les séquences de consommation réelles sur 7 jours, séquences qui sont nécessaires pour l'apprentissage et la recommandation de repas et non d'aliments isolés.
Un inconvénient de l'enquête INCA2 est la faible quantité de données disponibles. Il existe également l'enquête INCA3 plus récente : les recueils alimentaires de cette dernière portent sur 3 jours non-consecutifs (*versus* 7 jours pour INCA2), elle n'a pu être prise en compte dans le présent rapport mais son intégration est à l'étude pour la suite du projet EXERSYS.

Les données de consommation de l'enquête INCA2 sont interprétées dans le projet comme des préférences implicites des utilisateurs : on pourrait par exemple représenter les préférences implicites par une matrice d'interactions "aliments"/"utilisateurs" constituées de 0 (non consommation de l'aliment par l'utilisateur) et de 1 (consommation de l'aliment par l'utilisateur).

Il existe d'autres sources de données de préférences des utilisateurs tels que allrecipes, Marmiton où des recettes sont notées par des utilisateurs (le média Marmiton revendique 17 millions de visiteurs uniques par mois sur son site et 5 millions d'abonnés sur ses réseaux sociaux [2], le nombre de recettes n'est pas communiqué). Ces données massives ne sont malheureusement pas structurées comme les données de l'enquête INCA2 et ne portent pas sur des repas mais des aliments isolés. Ces sources pourraient être utilisées dans un second temps du projet pour enrichir les données disponibles sur les aliments et les préférences utilisateurs.

2.2.2 Données relatives à la nutrition

Les connaissances expertes en nutrition qui ont été utilisées dans le projet sont de différents types. Il s'agit notamment de :

- la définition des 44 groupes d'aliments dans l'enquête INCA2, présentés au tableau 2 : ils seront dénommés "groupes INCA2" dans le rapport
- la connaissance des comportements alimentaires des individus, provenant d'Experts en Nutrition sollicités tout au long du projet
- la composition nutritionnelle des aliments et des ingrédients de l'enquête INCA2, donnée par la table de composition CIQUAL [3].

L'un des enjeux du projet est de formaliser ces connaissances à l'aide des graphes de connaissance pour établir des règles et des contraintes nutritionnelles.

2.3 Intérêt du travail

2.3.1 Intérêt pour le domaine applicatif

L'adhésion de la population aux recommandations nutritionnelles est un défi majeur en termes de santé publique. Une alimentation saine aide à se protéger contre les maladies non transmissibles notamment le diabète, les maladies cardiaques, les accidents vasculaires cérébraux et le cancer (World Health Organization 2023). L'augmentation de la production d'aliments transformés, l'urbanisation et l'évolution des modes de vie ont entraîné une modification des habitudes alimentaires. La population consomme plus d'aliments riches en énergie, en graisses, en sucres libres et en sel/sodium qu'auparavant, et pour beaucoup d'individus insuffisament de fruits, légumes et autres fibres alimentaires comme les grains entiers.

2. https ://www.reworldmediaconnect.com/nos-marques/art-de-vivre/marmiton/
3. https ://ciqual.anses.fr/

Code Groupe aliments INCA	Nom Groupe d'aliments INCA
1	pain et panification sèche
2	céréales pour petit déjeuner
3	pâtes
4	riz et blé dur ou concassé
5	autres céréales
6	viennoiserie
7	biscuits sucrés ou salés et barres
8	pâtisseries et gâteaux
9	lait
10	ultra-frais laitier
11	fromages
12	oeufs et dérivés
13	beurre
14	huile
15	margarine
16	autres graisses
17	viande
18	volaille et gibier
19	abats
20	charcuterie
21	poissons
22	crustacés et mollusques
23	légumes (hors pommes de terre)
24	pommes de terre et apparentés
25	légumes secs
26	fruits
27	fruits secs et graines oléagineuses
28	glaces et desserts glacés
29	chocolat
30	sucres et dérivés
31	eaux
32	boissons fraîches sans alcool
33	boissons alcoolisées
34	café
35	autres boissons chaudes
36	pizzas, quiches et pâtisseries salées
37	sandwichs, casse-croûte
38	soupes et bouillons
39	plats composés
41	entremets, crèmes desserts et laits gélifiés
42	compotes et fruits cuits
43	condiments et sauces
44	aliments destinés à une alimentation particulière
45	-

FIGURE 2 – *Nomenclature des 44 groupes d'aliments dans l'enquête INCA2*

Le régime alimentaire est influencé par de nombreux facteurs sociaux et économiques qui interagissent de manière complexe pour façonner les habitudes alimentaires individuelles. Ces facteurs comprennent le revenu, les prix des denrées alimentaires (qui affecteront la disponibilité et l'abordabilité d'aliments sains), les préférences et les croyances individuelles, les traditions culturelles et les aspects géographiques et environnementaux.
Par conséquent, la promotion d'un environnement alimentaire sain - y compris des systèmes alimentaires qui favorisent une alimentation diversifiée, équilibrée et saine - nécessite la participation de multiples secteurs et parties prenantes, y compris le gouvernement et les secteurs public et privé (World Health Organization 2023).

Un système de recommandation qui prend en compte à la fois les préférences des utilisateurs et les contraintes nutritionnelles peut jouer un rôle déterminant pour la promotion de bonnes habitudes alimentaires et l'adhésion aux recommandations nutritionnelles.
L'intégration de règles à la recommandation alimentaire grâce aux graphes de connaissances permet non seulement d'intégrer des contraintes nutritionnelles mais aussi de prendre en compte des contraintes spécifiques des utilisateurs (e.g leurs allergies, le désir de ne manger que de la nourriture végétarienne, des contraintes liées à l'état de santé pour les sujets diabétiques...).

2.3.2 Etat de l'art tous domaines confondus

- **Systèmes de recommandation basés sur l'apprentissage**

  Les systèmes de recommandations ont été largement utilisés dans le domaines des films, livres, musique, achats en ligne qui sont des applications classiques (Delporte, Canu et Karatzoglou 2014).
  Les objectifs des systèmes de recommandation peuvent être divers. Il s'agit notamment pour les fournisseurs de ces systèmes d'augmenter le nombre d'articles vendus, fidéliser les consommateurs, mieux comprendre ce qu'ils souhaitent. Pour les utilisateurs, les objectifs sont également variés (trouver les meilleurs articles, influencer...) et il est clé pour tout système de recommandation de définir ses objectifs en amont, à la fois pour le fournisseur et pour l'utilisateur (Ricci, Rokach et Shapira 2015).

  Le principe d'un système de recommandation consiste pour un utilisateur donné à prédire son score d'appétence ou d'utilité pour une liste de produits, ce qui permet de les ordonner. Ces scores sont personnalisés et chaque utilisateur dispose de ses propres recommandations. (Delporte, Canu et Karatzoglou 2014). Les systèmes de recommandation utilisent trois types de données : les produits, les utilisateurs et les relations entre produits et utilisateurs (transactions). Les systèmes de recommandation les plus utilisés sont en général assez pauvres en contenu. Certaines techniques intègrent des connaissances (descriptions ontologiques des utilisateurs ou des produits), des contraintes, les relations sociales des utilisateurs mais elles sont moins fréquemment utilisées (Ricci, Rokach et Shapira 2015) .
  Les appréciations des utilisateurs peuvent être explicites telles que des notes données aux produits utilisés ou consultés, ou implicites, telles que des clics ou achats en ligne, i.e. des signaux de comportement de l'utilisateur qui peuvent indiquer ses préférences.

  Historiquement, les systèmes de recommandation sont classés en trois groupes sur la base de l'approche utilisée pour générer les recommandations (Adomavicius et Tuzhilin 2005) : l'approche de filtrage basée sur le contenu, l'approche de filtrage collaboratif, l'approche hybride.
  Dans l'approche de filtrage basé sur le contenu, le système apprend à recommander des éléments similaires à ceux que l'utilisateur a aimés dans le passé. La similarité des éléments

est calculée en fonction des caractéristiques associées aux éléments comparés. L'approche de filtrage collaboratif fait des recommandations à un utilisateur sur la base d'éléments que d'autres utilisateurs ayant des goûts similaires ont aimés dans le passé. La similitude de goûts de deux utilisateurs est calculée en fonction de la similitude dans l'historique des évaluations des utilisateurs (Ricci, Rokach et Shapira 2015). Les deux grands modèles de filtrage collaboratif sont les méthodes basée sur le voisinage (orientées utilisateur ou orientées produit) et les modèles de facteurs latents ou factorisation matricielle (Koren, Bell et Volinsky 2009).

Les deux approches présentent des limites. Les systèmes basés sur le contenu ne tiennent généralement pas compte de la qualité des éléments dans le processus de recommandation (Lops, Jannach et Musto 2019). Les sytèmes basés sur le filtrage collaboratif souffrent du démarrage à froid, i.e. la difficulté à traiter les nouveaux produits et les nouveaux utilisateurs (Ricci, Rokach et Shapira 2015). L'approche hybride a pour avantage de pallier les limites et tirer parti de chaque système : à la fois des informations de préférence issues du filtrage collaboratif et d'informations contextuelles propres aux utilisateurs ou aux articles que fournissent les systèmes basés sur le contenu(Delporte, Canu et Karatzoglou 2014).

- **Graphes de connaissance et application à la recommandation**

  Un graphe de connaissance est un graphe hétérogène où les nœuds fonctionnent comme des entités et les arêtes représentent les relations entre les entités. Les objets et leurs attributs peuvent être cartographiés dans le graphe pour comprendre les relations mutuelles entre les éléments (Guo et al. 2020).
  Les graphes de connaissance ont été proposés pour enrichir la représentation des utilisateurs ou des produits dans l'objectif d'ameliorer la performance de la recommandation basée sur des méthodes classiques de type filtrage collaboratif (Zhang et al. 2016 ; Wang et al. 2019) mais cette approche est encore peu décrite dans la littérature.

2.3.3 Etat de l'art dans le domaine de la nutrition

Dans le domaine la nutrition, on peut définir différents besoins auxquels doit répondre un système de recommandation afin d'assurer la satisfaction des utilisateurs :

— **BP :** le besoin de personnalisation de la recommandation

— **BN :** le besoin de prise en compte des contraintes nutritionnelles spécifiques d'un utilisateur (ses allergies, ses préférences...)

— **BS :** le besoin de séquentialité des aliments et de cohérence du repas

- **Systèmes de recommandation**

  Dans le domaine de l'alimentation, les systèmes de recommandation sont principalement basés sur le filtrage collaboratif, appliqué à la recommandation de recettes et non de repas (Trattner et Elsweiler 2017).
  Ceci est une limite à leur utilisation dans la vie réelle puisque les aliments sont consommés de façon complémentaire si l'on considère par exemple une structure de repas recommandé pour un adulte, repas qui comporte entrée, plat principal avec protéines animales, légumes et féculents, dessert (Santé publique France s. d.).

- **Graphes de connaissance**

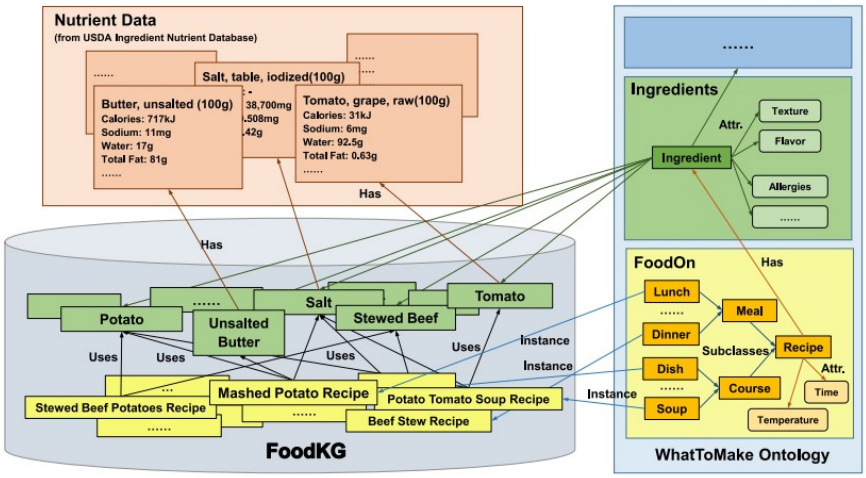

FIGURE 3 – *FoodKG, exemple de graphes de connaissance en alimentation d'après* MIN *et al. 2022*

Il existe des graphes de connaissance appliqués au domaine de l'alimentation mais ils sont spécialisés dans des domaines spécifiques tels que l'agriculture, la production, des états de santé spécifiques. Il y a peu de graphes de connaissances orientés vers la population générale et les consommateurs qui souhaitent manger sainement. FoodKG en est un exemple : ce système s'appuie sur trois sources de données : des recettes, le contenu nutritionnel des ingrédients (source : United States Department of Agriculture) et une ontologie alimentaire pour organiser les ingrédients (cf. Figure 3) (HAUSSMANN et al. 2019).

En conclusion, dans le domaine de l'alimentation, aucun système de recommandation combinant les deux méthodes (recommandation par apprentissage automatique des préférences consommateurs et ingrégration d'informations nutritionnelles par graphes des connaissances) n'a pu être identifié, ce qui illustre le caractère novateur du projet.

2.3.4 Caractère novateur du projet

La combinaison de deux méthodes pour générer une recommandation (apprentissage des préférences alimentaires des utilisateurs et post-filtrage suivant des règles à l'aide des graphes de connaissance) permet :

— de prendre en compte à la fois **les préférences des utilisateurs et l'équilibre nutritionnel des recommandations**, tandis que la plupart des systèmes de recommandation alimentaires existants prennent en compte un seul aspect (TRATTNER et ELSWEILER 2017 ; YERA TOLEDO, ALZAHRANI et MARTÍNEZ 2019 ; SILVA et al. 2022 ; PECUNE, CALLEBERT et MARSELLA 2020)

— de **compenser les faiblesses d'une méthode** grâce à l'autre. Les systèmes de recommandation de type filtrage collaboratif souffrent du démarrage à froid : la non disponibilité des informations nécessaires pour faire une recommandation à de nouveaux utilisateurs (GOPE et JAIN 2017). Grâce aux graphes de connaissances, il est possible de définir des profils utilisateurs (basés sur leurs préférences ou habitudes alimentaires) et de rattacher un nouvel utilisateur à un profil prédéfini après lui avoir posé quelques questions concises

— **d'éviter les recommandations non pertinentes** pour un utilisateur (e.g. proposer de la viande à un consommateur végétarien), recommandations qui pourraient discréditer le système et limiter l'adhésion des utilisateurs.

Au-delà du projet, la combinaison des 2 méthodes permettra également :

— de considérer la **diversité des repas recommandés**, en raisonnant à l'échelle d'un jour ou d'une semaine

— d'intégrer des **éléments de contexte** du repas, tels que le lieu (domicile, restaurant), le moment et le contexte social du repas qui sont de forts déterminants des goûts et des consommations(Trattner et Elsweiler 2017)

Une approche innovante du projet consiste aussi à appliquer des algorithmes existants au domaine alimentaire, à savoir :

— **l'algorithme Word2Vec**, un algorithme largement employé pour la recommandation de différents produits (Caselles-Dupré, Lesaint et Royo-Letelier 2018) mais pas pour la recommandation alimentaire,

— **les réseaux de neurones récurrents (RNN)**, modèle certes utilisé pour la recommandation basée sur la session utilisateur (Hidasi et al. 2015), pour la prédiction des prochains aliments susceptibles d'être achetés en ligne sur la base de l'historique (Lee et al. 2020), mais pas pour la recommandation de repas.

L'utilisation de ces algorithmes dans nos modèles sera expliquée aux parties 3 et 4.

3 Modèle FiltreCollab (axe 1) : recommandation de type filtrage collaboratif

3.1 Préambule

L'objectif de l'axe 1 est d'élaborer une recommandation de repas basée sur les préférences alimentaires des utilisateurs par apprentissage automatique, tout en tenant compte des connaissances experts sur la structure "type" d'un repas.

Notre premier modèle présenté ci-après dans la section 3 est une approche de type filtrage collaboratif, il est dénommé ci-après **FiltreCollab**. Il est basé sur l'apprentissage d'un espace de représentation des aliments et la modélisation des utilisateurs. Ce modèle qui constitue la majeure partie de mon stage et du présent rapport a été proposé, paramétré et testé sur le petit déjeuner (repas qui présente une moindre variété en termes d'aliments) mais il a montré des limites sur le déjeuner/diner qui sont des repas plus complexes.

Un deuxième modèle, appelé **GenSeqRNN** (génération de séquence par réseau de neurones récurrents), a été paramétré et testé sur le déjeuner/diner pour exploiter le caractère séquentiel de ces repas et améliorer la performance. Il sera présenté à la section 4.

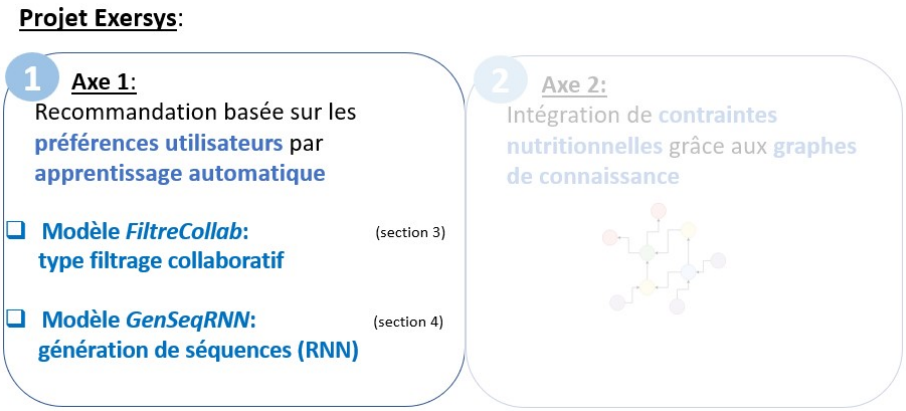

FIGURE 4 – *Deux modèles testés dans l'axe 1 du projet EXERSYS*

3.2 Méthode pour le modèle FiltreCollab

3.2.1 Approche globale

Dans ce premier modèle **FiltreCollab**, nous souhaitons apprendre un espace de représentation des aliments dans lequel les distances entre aliments ont un sens (étape 1), définir des catégories d'aliments substituables (étape 2), au sein de ces catégories définir des probabilités de tirage des aliments (étape 3), puis définir des modalités de tirage pour les différentes catégories afin d'émettre une recommandation de repas pour un utilisateur (étape 4).

— **Etape (1)** : avec l'utilisation de l'algorithme Word2Vec entrainé sur les séquences des repas des individus, nous faisons l'hypothèse que des aliments consommés dans un même contexte sont proches dans l'espace et sont "substituables". Une approche alternative aurait été l'apprentissage des préférences utilisateurs par des méthodes de filtrage collaboratif type factorisation matricielle pour rechercher des communautés de goûts, mais ce n'est pas celle que nous avons retenue ici.

— **Etape (2)** : nous cherchons à constituer des catégories d'aliments substituables.

— **Etape (3)** : à partir de ces catégories nous définissons les probabilité de tirage des aliments au sein d'une catégorie, probabilités basées sur les distances (plus exactement sur les similarités) entre utilisateurs et aliments au sein d'une catégorie. Nous faisons l'hypothèse que les distances apprises lors de l'apprentissage de l'espace de représentation des aliments ont du sens au sein de chaque catégorie définie.

— **Etape (4)** : nous définissons les modalités de tirage des différentes catégories et leurs associations/exclusions pour constituer une recommandation de repas cohérent.

L'approche globale est décrite à la Figure 5.

Nous avons appliqué l'approche d'abord au petit déjeuner et dans un second temps aux repas plus complexes que sont le déjeuner et le dîner.
Nous définissons des critères d'évaluation à chaque étape du processus pour valider les choix des hyperparamètres de notre modèle et évaluer sa performance par rapport à des systèmes de recommandation de référence.

Pour implémenter notre modèle **FiltreCollab**, le langage utilisé est python et les méthodes sont celles de l'apprentissage automatique avec des algorithmes utilisés dans les modèles de langue, les systèmes de recommandations, la classification non supervisée (bibliothèques gensim, scikit-learn,

lightfm).

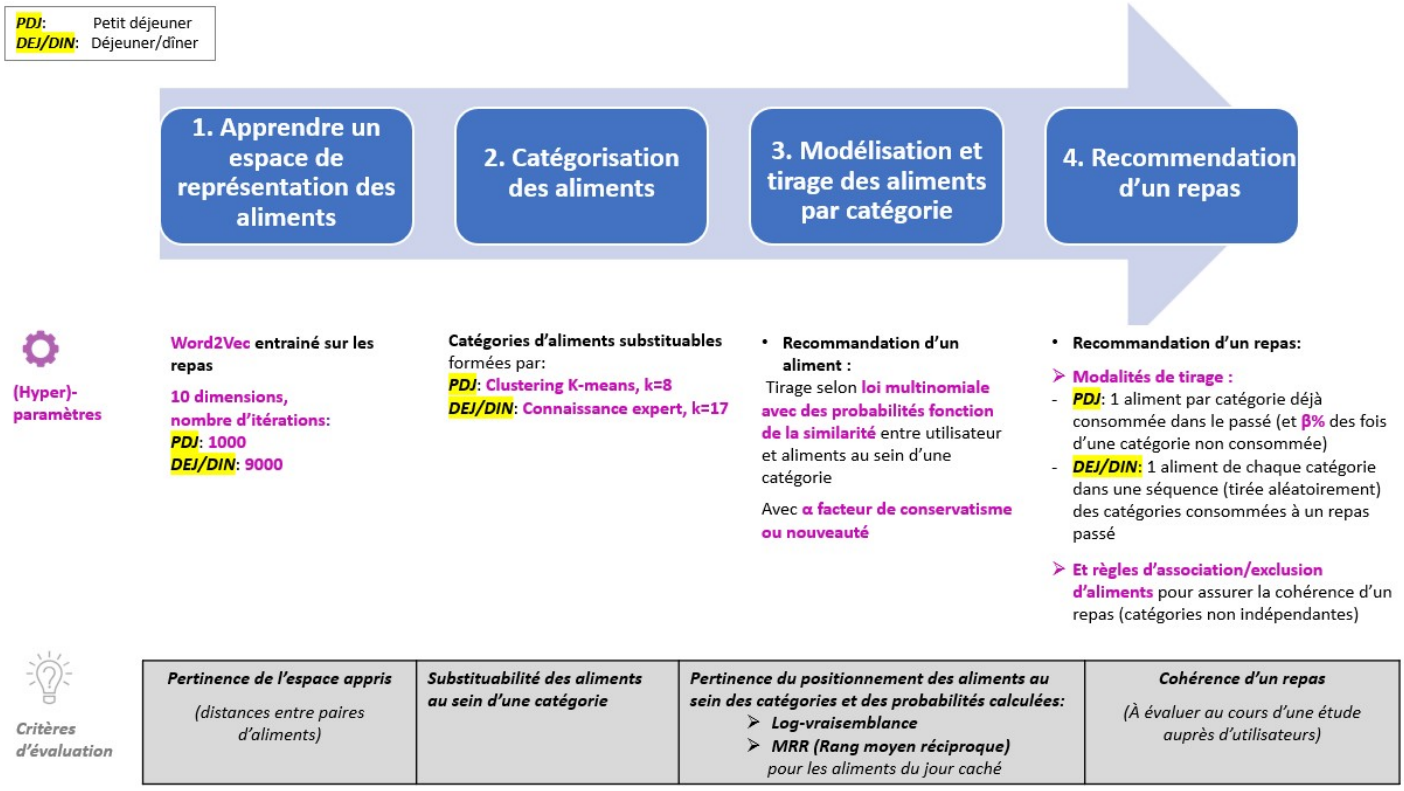

FIGURE 5 – *Paramètres du modèle et critères d'évaluation définis à chaque étape du modèle FiltreCollab (valables pour petit déjeuner et déjeuner/dîner si non spécifiés)*

3.2.2 Etape 1 : Apprentissage de l'espace de représentation des aliments où les aliments consommés dans un contexte similaire sont proches

L'algorithme utilisé est **Word2vec** (W2V) (MIKOLOV, SUTSKEVER et al. 2013 ; MIKOLOV, K. CHEN et al. 2013). Il s'agit d'un algorithme d'apprentissage automatique auto-supervisé qui permet de représenter les mots sous forme de vecteurs dans un espace dont le nombre de dimensions est défini par l'utilisateur. Cette technique de traitement automatique du langage naturel est utilisée pour l'analyse sémantique et la classification de texte.
L'algorithme Word2vec apprend des représentations vectorielles des mots à partir d'un texte et permet de prédire le contexte dans lequel chaque mot apparaît dans le texte. Les représentations vectorielles de mots ainsi ont la propriété intéressante que les mots qui apparaissent dans le même contexte ont des vecteurs "proches" dans l'espace et sont caractérisés par une similarité élevée. La mesure de la similarité entre 2 mots est la similarité cosinus entre leurs 2 vecteurs, comprise entre -1 et 1. La bibliothèque Python utilisée est gensim.

L'application de **Word2vec** à des séquences d'aliments réellement consommés au sein d'un repas par les individus de l'étude permet d'identifier au sein d'un repas des aliments similaires car consommés dans un même contexte ("entourés" des mêmes aliments) et substituables. Après exclusion des aliments rares (consommés par moins de 5% des individus), on utilise l'algorithme **Word2vec** pour appendre les petits déjeuners (constitués des 55 aliments non rares) ou les déjeu-

ners et diners (constitués de 335 aliments non rares) pour les individus de l'étude INCA2, chaque séquence étant un repas de la semaine consommé par un individu.

Les principaux paramètres à fixer sont le nombre n de dimensions de l'espace de représentation (fixé à 10, l'ordre de grandeur de la taille des plus petites catégories), le nombre d'itérations pour l'apprentissage (1000 pour le petit déjeuner, 9000 pour le déjeuner/diner) et la taille de la fenêtre (que l'on fixe à la longueur maximale d'un repas, aliments rares exclus). Chaque aliment est représenté par un vecteur à n dimensions dans l'espace appris. Le choix du nombre de dimensions est un critère majeur car un nombre trop élevé peut conduire à un surajustement mais il y a peu de publications consacrées au sujet (Hung et Yamanishi 2021).

Nous définissions un **critère d'évaluation pour cette étape** : la pertinence de l'espace appris en considérant les distances entre paires d'aliments (au regard de la connaissance expert, i.e. le groupe INCA2 auquel ces aliments appartiennent).

3.2.3 Etape 2 : regroupement des aliments en catégories d'aliments substituables

Nous cherchons à obtenir des catégories d'aliments substituables au sein d'un repas. Nous avons utilisé deux méthodes :

— le clustering par algorithme K-means

— la constitution de catégories d'aliments substituables par la connaissance expert (en cas d'échec par clustering)

K-means est l'une des méthodes de clustering non supervisées les plus utilisées, rapides et robustes. C'est une méthode de partitionnement qui regroupe les données en maximisant la similarité entre objets appartenant au même groupe tout en minimisant la similarité entre objets appartenant à différents groupes. La bibliothèque Python utilisée est scikit-learn.
Le principal paramètre à définir est le nombre de groupes k, qui a été principalement déterminé grâce à la connaissance expert (homogénéité des catégories par rapport aux groupes d'aliments INCA2) et dans une moindre mesure à l'aide du score de silhouette [4] qui est une mesure de la qualité du clustering.

Nous avons pu obtenir des catégories homogènes par clustering avec K-means sur le petit déjeuner, mais pas dans le cas des déjeuners/diners où le clustering conduisait à des catégories hétérogènes et regroupant des aliments non substituables. Cela explique le recours à la seconde méthode : l'utilisation des connaissances experts *a priori* pour constituer des catégories dans le cas des déjeuners et diners (catégories correspondant à un ou plusieurs groupes INCA2).

Nous définissions un **critère d'évaluation pour cette étape** : le degré d'homogénéité des catégories par rapport aux groupes INCA2 et la substituabilité des aliments au sein d'une catégorie, évaluée par l'expert.
Il est à noter que l'appartenance d'aliments à un même groupe INCA2 assure leur substituabilité. Mais des aliments appartenant à des groupes INCA2 différents peuvent être jugés substituables (ex. thé et café).

3.2.4 Etape 3 : modélisation des catégories/utilisateurs et probabilités de tirage des aliments au sein des catégories

L'approche est identique pour le petit déjeuner et le déjeuner/dîner.

4. Le coefficient de silhouette est calculé en utilisant la distance intra-groupe moyenne, a, et la distance moyenne au groupe le plus proche, b, pour chaque échantillon. Le coefficient de silhouette pour un échantillon est $\frac{b-a}{max(b-a)}$.

On choisit de représenter une catégorie c (c appartenant à C l'ensemble des catégories) par la moyenne m_c des aliments qui figurent dans cette catégorie.
On choisit de modéliser un utilisateur u :

— par ses fréquences de consommation f_c de chaque catégorie c d'aliments (nombre d'aliments de la catégorie c divisé par le nombre total d'aliments consommés aux petits déjeuners des jours de recueil),

— au sein de chaque catégorie c, par la moyenne u_c des aliments de la catégorie c qu'il a consommés aux repas au cours des jours de recueil (u_c : moyenne de l'utilisateur u pour la catégorie c)

— pour chaque catégorie c, pour chaque aliment k au sein de la catégorie c, par la probabilité $P_{k|c,u}$ que l'aliment k soit recommandé à l'utilisateur u par tirage aléatoire dans la catégorie c selon une loi multinomiale .

Cette probabilité est définie à l'aide d'une fonction softmax :

— soit en fonction de la distance entre l'aliment et l'utilisateur (u_c, la moyenne des aliments consommés au sein de la catégorie) s'il a consommé la catégorie ($f_c \neq 0$),

— soit en fonction de la distance entre l'aliment et la moyenne des aliments de la catégorie (m_c) si l'utilisateur n'a pas consommé d'aliments dans cette catégorie ($f_c = 0$).

Avec l'introduction d'un facteur α, nous ajoutons la possibilité de réduire ou augmenter les écarts de probabilité entre aliments pour favoriser le conservatisme (en augmentant α) ou au contraire la nouveauté des aliments recommandés (en diminuant α).

Soit $P_{k|c,u}$ la probabilité de recommander un aliment k de la catégorie c à un utilisateur u, soit I_c l'ensemble des aliments de la catégorie c et i un aliment appartenant à I_c, S_{i,u_c} la similarité entre l'utilisateur représenté par u_c sa moyenne au sein de la catégorie et l'aliment i, S_{i,m_c} la similarité entre m_c la moyenne de la catégorie et l'aliment i, et α le facteur favorisant le conservatisme ou la nouveauté des aliments recommandés. La probabilité $P_{k|c,u}$ est ainsi définie ci-dessous :
Si $f_c \neq 0$:

$$P_{k|c,u} = \frac{e^{\alpha * S_{k,u_c}}}{\sum_{i \in I_c} e^{\alpha * S_{i,u_c}}} \tag{1}$$

Si $f_c = 0$:

$$P_{k|c,u} = \frac{e^{\alpha * S_{k,m_c}}}{\sum_{i \in I_c} e^{\alpha * S_{i,m_c}}} \tag{2}$$

S_{i,u_c} et S_{i,m_c} sont des similarités cosinus, égales au cosinus de l'angle de deux vecteurs. La similarité cosinus est fréquemment utilisée pour comparer des mots dans le traitement automatique du langage et mesurer leur proximité sémantique, avec l'algorithme **Word2vec** notamment (Mikolov, Sutskever et al. 2013).

S_{i,u_c} est ainsi calculée de la manière suivante :

$$S_{i,u_c} = \frac{u_c \cdot i}{\|u_c\| \|i\|} \tag{3}$$

avec $u_c \cdot i$ produit scalaire des vecteurs u_c et i, $\|u_c\|$ norme du vecteur u_c, $\|i\|$ norme du vecteur i

Nous sommes ainsi capables de recommander un aliment au sein de chaque catégorie par tirage aléatoire suivant une loi multinomiale selon les probabilités définies ci-dessus.

Nous définissions un **critère d'évaluation à cette étape** pour évaluer la pertinence du positionnement des aliments et des probabilités calculées à partir des distances (entre aliments/utilisateurs) au sein de chaque catégorie.
Il existe différentes métriques de calcul de la performance d'une recommandation [5] ; nous avons retenu le score de **log-vraisemblance des consommations du jour caché** car notre modèle est probabiliste, ainsi que le **MRR ("Mean Reciprocal Rank")** car il s'agit d'une métrique de rang communément utilisée dans le domaine de la recommandation.
Pour calculer ces deux métriques, on divise le jeu de données en 2 ensembles : un jeu d'entrainement qui est constitué des repas (petits-déjeuners ou déjeuner/diner) des 6 premiers jours de recueil et un jeu de test qui est le repas(petit déjeuner ou déjeuner/diner) du dernier jour. Le jour caché est un jour aléatoire de la semaine puisque les individus de l'enquête ont pu débuter l'enquête n'importe quel jour de la semaine. On mesurera la pertinence de notre recommandation (qui est basée sur le jeu d'entrainement) selon sa capacité à "prédire" le jour caché.

Les deux métriques sont définies ci-dessous :

— On calcule la moyenne sur l'ensemble des individus du score de log-vraisemblance selon notre modèle :
En considérant $\mathcal{L}_u$ la log-vraisemblance des consommations du jour caché pour un utilisateur u, I_{uc} l'ensemble des aliments de la catégorie c consommés par l'utilisateur u le jour caché, C l'ensemble des catégories d'aliments du petit déjeuner, $P_{c,u}$ la probabilité pour l'utilisateur u de consommer la catégorie c et $P_{i|c,u}$ la probabilité pour l'utilisateur u de consommer l'aliment i sachant la catégorie c, on a la formule suivante :

$$\mathcal{L}_u = \log\left(\prod_{c\in C}\prod_{i\in I_{uc}} P_{c,u}\cdot P_{i|c,u}\right) \tag{4}$$

Avec :
si $f_c \neq 0$: $P_{i|c,u}$ définie à l'équation (1) et $P_{c,u} = f_c * (1 - |C| * \beta) + \beta$
si $f_c = 0$: $P_{i|c,u}$ définie à l'équation (2) et $P_{c,u} = \beta$
β : hyperparamètre du modèle, modélisant la probabilité pour un individu de consommer une catégorie non encore consommée les 6 jours précédents (fixé arbitrairement à une valeur faible initialement)

— On calcule la moyenne sur l'ensemble des individus du MRR selon notre modèle : En considérant MRR_u le MRR d'un utilisateur u, N le nombre total d'aliments consommés par l'utilisateur au petit déjeuner le jour caché, I_{uc} l'ensemble des aliments de la catégorie c consommés par l'utilisateur u, $rang\ i_c$ le rang du i^{eme} aliment consommé par l'utilisateur le jour caché au sein de la recommandation faite par le système pour la catégorie c, on a la formule suivante :

$$MRR_u = \frac{1}{N} * \left(\sum_{i\in I_{uc}} \frac{1}{rang\ i_c}\right) \tag{5}$$

5. Différentes métriques ont été appliquées pour mesurer la performance des systèmes de recommandation dans la littérature. Ceux-ci reflètent généralement l'erreur des notes prédites (par exemple MAE "Mean Absolute Error" "ou RMSE "Root Mean Square Error") ou la qualité de la liste des n premiers éléments classés, par exemple le Rang moyen réciproque (MRR "Mean Reciprocal Rank"), le rappel, la précision, précision moyenne (MAP "Mean Average Precision") et le gain cumulé actualisé normalisé (NDCG "Normalized Discounted Cumulative Gain") (Trattner et Elsweiler 2017 ; M. Chen et Liu 2017).

Les scores moyens des 2 métriques sont comparés aux valeurs trouvées pour des modèles de référence : recommandation aléatoire par catégorie, recommandation des aliments les plus populaires. Dans le cas de la métrique de MRR (rang moyen récipoque), le score de notre modèle est comparé avec un modèle de recommandation BPR ("Bayesian Personalized Ranking") qui est une méthode de factorisation matricielle (Rendle et al. 2009).
Notre démarche est proche de celle adoptée par Trattner et Elsweiler 2017 pour le choix des systèmes de recommandation alimentaires de référence.

3.2.5 Etape 4 : Modalités de recommandation d'un repas à un utilisateur à partir des catégories

Pour la recommandation d'un repas pour un utilisateur donné, plusieurs approches sont possibles :

— tirer un aliment par catégorie déjà consommée dans le passé et dans $\beta\%$ des cas par catégorie non encore consommée (appproche favorisée pour le petit déjeuner)

— tirer aléatoirement une séquence de catégories consommée lors d'un repas des jours précédents, pour tirer un aliment au sein de chacune de ces catégories (approche favorisée pour le déjeuner et diner)

Le tirage d'un seul aliment au sein d'une catégorie est justifié par les fréquences de consommations observées : zéro ou un aliment de chaque catégorie consommé par au moins 98% des individus pour 7 des 8 catégories dans le cas du petit déjeuner par exemple.

Pour les tirages, on considère que les catégories sont indépendantes les unes des autres, c'est à dire que la consommation d'une catégorie n'affecte pas les probabilités de consommation d'autres catégories (on se base sur l'historique de consommation). Pour corriger cette approximation, on tient ensuite compte des associations d'aliments existantes au sein d'un repas en définissant des règles d'association (exemple consommation de pain si beurre) ou d'exclusion de certaintes catégories en fonction de leurs indices de support, confiance et lift.

Nous définissons un **critère d'évaluation** pour la performance de recommandation d'un repas. Le critère usuel est la satisfaction de l'utilisateur suite à la recommandation, que nous ne pouvons évaluer à ce stade du projet.
Toutefois, une composante majeure de la satisfaction utilisateur nous paraît être la cohérence globale du repas recommandé. Nous considérons donc cette dernière comme critère d'évaluation à l'étape 4, et nous souhaitons l'évaluer ultérieurement par le biais d'une étude avec des utilisateurs (en soumettant des recommandations issues du sytème à un petit nombre d'utilisateurs recrutés pour qu'ils en évaluent la cohérence par une note par exemple, protocole à définir).

3.3 Résultats et discussion pour le modèle FiltreCollab

Les résultats sont résumés à la Figure 6 et détaillés ci-dessous.

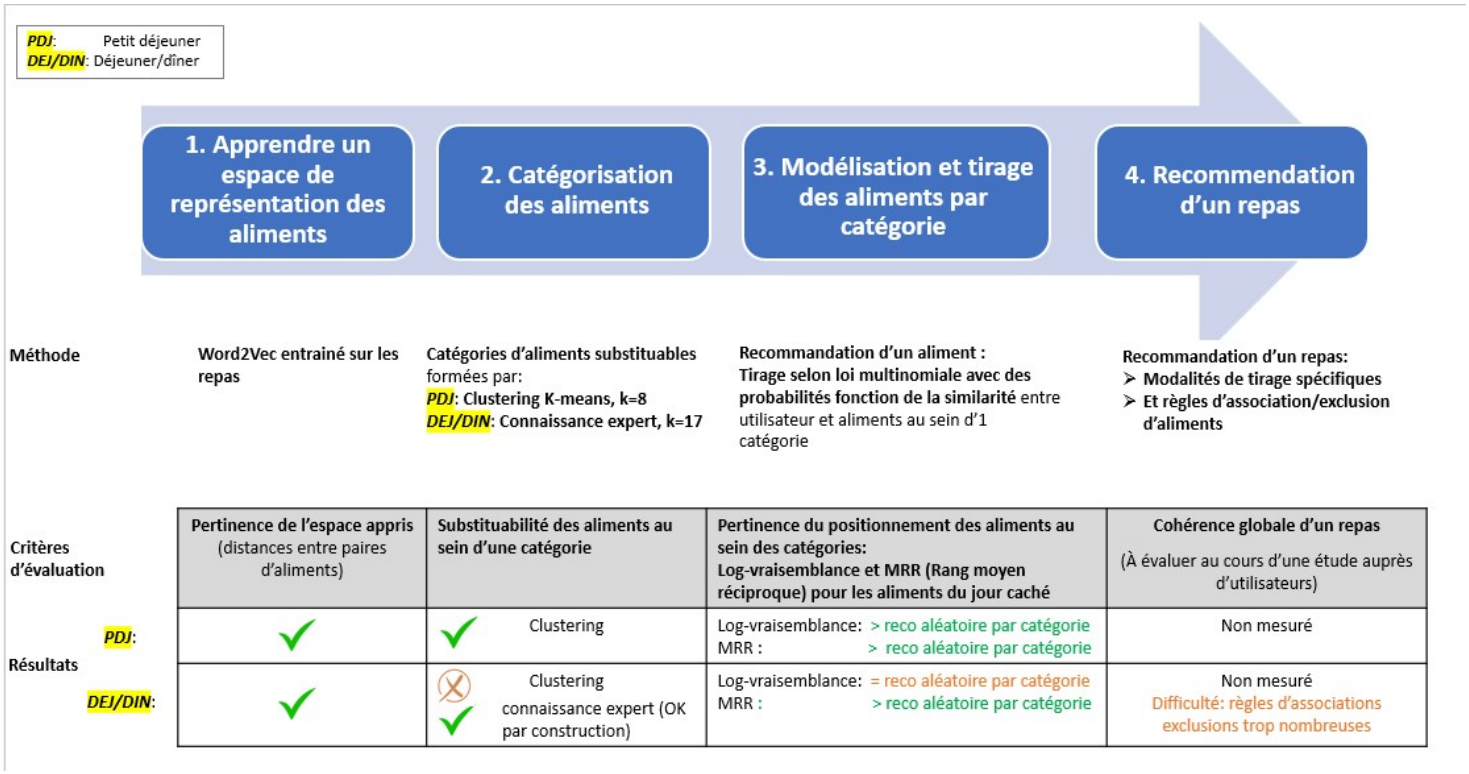

FIGURE 6 – *Critères d'évaluation et résultats à chaque étape du modèle FiltreCollab : comparaison petit déjeuner et déjeuner/diner)*

3.3.1 Etapes 1 et 2 : Apprentissage de l'espace de représentation des aliments et regroupement en catégories d'aliments substituables

Dans le cas du petit déjeuner :

— les distances entre paires d'aliments montrent la pertinence de l'espace appris après entrainement avec W2V (distances interprétées avec la connaissance expert)

— le clustering par K-means associé à la connaissance expert permet d'obtenir 8 catégories d'aliments, dont la majorité sont des catégories d'aliments substituables présentant une forte homogénéité par rapport aux groupes d'aliments INCA2

Ces catégories sont présentées à la Figure 7, et le degré d'homogénéité des catégories par rapport aux groupes INCA2 est illustré dans la deuxième colonne.

N° de catégorie constituée par clustering (k-means) [1]	Groupes INCA[2] et pourcentages des aliments qui appartiennent à ces groupes dans la catégorie	Noms des aliments de la catégorie
0	**lait: 17%, eaux: 17%, boissons fraîches sans alcool: 33%, fruits: 33%**	['eau du robinet', 'eau de source', 'lait demi-écrémé uht', 'orange fraîche', "jus d'orange à base de concentré pasteurisé", 'pomme non pelée fraîche', 'lait écrémé uht', 'banane fraîche', "jus d'orange pressé maison", "pur jus d'orange pasteurisé", 'jus de fruits sans précision', 'kiwi
1	**pain et panification sèche: 100%**	['pain baguette', 'pain de campagne ou bis', 'pain courant français boule à la levure', 'pain complet ou intégral artisanal', 'biscotte classique type heudebert lu', 'pain grillé maison', 'pain de mie', 'autre biscotte', 'pain aux céréales artisanal']
2	**beurre: 50%, margarine: 50%**	['matière grasse allégée 60% m.g.', 'matière grasse allégée 55-60% m.g. riche en oméga 3 et 6', 'beurre doux', 'beurre demi-sel sel maxi 3%', 'matière grasse légère 38-41% m.g. à tartiner', 'beurre allégé sans précision']
3	**café: 67%, autres boissons chaudes: 33%**	['café au lait ou café crème ou cappuccino non sucré', 'café noir prêt à boire non sucré', 'thé infusé non sucré', 'café soluble reconstitué prêt à boire non sucré', 'poudre cacaotée et sucrée pour boisson au chocolat', 'café expresso non sucré']
4	**viennoiserie: 67%, chocolat: 11%, biscuits sucrés ou salés et barres: 11%, pâtisseries et gâteaux: 11%**	['brioche industrielle préemballée', 'croissant sans précision', 'pain au chocolat feuilleté artisanal', 'brioche sans précision', 'croissant au beurre artisanal', 'pâte à tartiner au chocolat et aux noisettes type nutella', 'madeleine', 'pain au lait artisanal', 'goûter sec fourré au chocolat type prince ou bn au chocolat']
5	**sucres et dérivés: 75%, aliments destinés à une alimentation particulière: 25%**	['sucre blanc', 'sucre blanc ajouté au service', "édulcorant à l'aspartame", 'sucre roux', 'aliment non codifié']
6	**sucres et dérivés: 71%, ultra-frais laitier: 14%, fruits: 14%**	['confiture ou marmelade tout type', 'clémentine ou mandarine fraîche', 'yaourt ou spécialité laitière nature', 'miel', 'confiture de fraise', "confiture d'abricot", 'confiture allégée']
7	**autres boissons chaudes: 50%, café: 50%**	['poudre soluble à base de chicorée et de café type ricoré', 'café soluble en poudre']

FIGURE 7 – *Composition des 8 catégories d'aliments du petit-déjeuner issues du clustering avec K-means, k principalement déterminé par connaissance expert (modèle FiltreCollab)*

Dans le cas du déjeuner et du dîner :

— les distances entre paires d'aliments montrent la pertinence de l'espace appris après entrainement avec W2V (distances interprétées avec la connaissance expert)

— mais le clustering par K-means, quel que soit le nombre de catégories k, conduit à l'obtention de certaines catégories hétérogènes par rapport aux groupes d'aliments INCA2 ou à la connaissance expert et qui ne sont pas constituées d'aliments substituables

Dans le cas du déjeuner et du dîner, un bruit important et un nombre limité de données pour l'apprentissage des séquences de repas avec W2V n'ont pas permis de positionner correctement tous les aliments dans l'espace et de rapprocher tous les aliments substituables. Ces résultats peuvent être liés à :

— la variété plus importante des aliments consommés et le nombre de combinaisons possibles très supérieur au déjeuner/diner par rapport au petit déjeuner

— l'existence de moments de consommation qui structurent le déjeuner et le diner (entrée, plats, dessert) i.e. une séquentialité des aliments consommés, qui n'est pas prise en compte dans l'apprentissage (ce qui n'est pas le cas pour le petit déjeuner).

Nous avons décidé d'intégrer de la connaissance expert pour tenter de compenser ces limitations, et avons constitué 16 catégories d'aliments substituables sur la base des groupes INCA2 et de l'étude des motifs fréquents. Nous avons ainsi regroupé au sein d'une même catégorie des groupes INCA2 d'aliments qui sont substituables d'un point de vue nutritionnel et qui ne sont pas consommés ensemble lors d'un repas : citons par exemple les 4 groupes INCA2 "viande", "volaille et gibier", "poissons", "oeufs et dérivés" qui sont regroupés dans la même catégorie). Les 16 catégories sont présentées au Tableau suivant.

N Catégorie constituée par connaissance expert[1]	Codes Groupes d'aliments INCA[2]	Noms Groupes d'aliments INCA[2]	Noms des aliments
0	36, 3, 4, 39, 24	pizzas, quiches et pâtisseries salées, pâtes, riz et blé dur ou concassé, plats composés, pommes de terre et apparentés	'pizza jambon fromage', 'pizza spéciale', 'pizza sans précision', 'pizza royale (jambon fromage champignon)', 'autre pizza', 'quiche lorraine', 'pâtes alimentaires cuites', 'pâtes alimentaires aux oeufs cuites', 'blé dur précuit nature type ebly', 'couscous (graine seule) semoule cuite', 'riz blanc cuit', 'blanquette de veau', 'cassoulet en conserve', 'choucroute garnie en conserve', 'pot-au-feu', 'paella', 'boeuf bourguignon', 'couscous royal à la viande', 'hachis parmentier', 'raviolis à la viande à la sauce tomate en conserve', 'gratin dauphinois', 'lasagnes à la bolognaise', 'salade de pommes de terre', 'crêpe fourrée béchamel jambon fromage', 'nem ou pâté impérial', 'nugget de volaille', 'cordon bleu de volaille type père dodu ou le gaulois', 'tomate farcie', 'tomate à la provençale', 'riz cantonais', 'taboulé ou salade de couscous', 'pomme de terre cuite au four', "pomme de terre cuite à l'eau", 'pomme de terre frite non salée', 'pomme de terre cuite à la vapeur', 'pomme de terre frite surgelée cuite', 'pomme noisette précuite surgelée', 'purée de pomme de terre avec lait et beurre non salée', 'purée de pomme de terre à base de flocons reconstituée au lait demi-écrémé', 'pomme de terre dauphine cuite', 'pomme de terre rissolée surgelée'
1	10	ultra-frais laitier	'crème fraîche ou crème de lait sans précision', 'crème chantilly sucrée sous pression uht', 'crème fraîche ou crème de lait fluide allégée 15 à 20% m.g. uht', 'crème fraîche allégée ajoutée au service', 'crème fraîche non allégée ajoutée au service', 'yaourt ou spécialité laitière nature', 'yaourt brassé nature au lait entier type danone velouté nature', 'yaourt nature 0% m.g. type taillefine nature 0% danone yoplait silhouette', 'yaourt nature au lait demi-écrémé sucré', 'yaourt aromatisé au lait entier sucré', 'yaourt ou spécialité laitière sans précision', 'yaourt ou spécialité laitière aux fruits sans précision', 'yaourt aromatisé au lait demi-écrémé sucré', 'yaourt aux fruits au lait entier sucré type danone recette crémeuse aux fruits', 'yaourt aux fruits au lait demi-écrémé sucré type panier de yoplait', 'spécialité laitière type yaourt 0% aux fruits sucrée type sveltesse aux fruits p', 'spécialité laitière type yaourt 0% aux fruits aux édulcorants type panier de yop', 'lait fermenté au bifidus nature au lait entier type danone activa', 'spécialité laitière nature à base de fromage blanc enrichie en crème type fjord', 'yaourt bulgare ou velouté à la pulpe de fruits au lait entier', 'fromage blanc sans précision', 'fromage blanc battu 20% m.g. nature'
2	43	condiments et sauces	'mayonnaise', 'sauce tomate sans viande', "vinaigrette à l'huile d'olive", 'vinaigrette allégée', 'vinaigrette sans précision', 'sauce béchamel en brique', 'sauce tomate à la viande ou bolognaise type panzani ou buitoni', 'sauce au beurre', 'sauce à la crème', 'sauce à la crème aux épices', 'sauce au vin rouge', 'vinaigrette ajoutée au service', 'vinaigrette allégée ajoutée au service', 'autre sauce', 'ciboule ou ciboulette fraîche', 'persil frais', 'basilic frais', 'herbes aromatiques fraîches', 'poivre noir moulu', 'sel fin ou gros', 'sel de mer', 'épices sans précision', 'cornichon au vinaigre', 'ketchup', 'moutarde', 'olive verte en saumure', 'ail frais', 'vinaigre', 'concentré de tomate'
3	17, 18, 21, 12	viande, volaille et gibier, poissons, oeufs et dérivés	'agneau côtelette grillée', 'agneau gigot rôti', 'boeuf entrecôte grillée', 'boeuf faux-filet grillé', 'boeuf bifteck grillé', 'boeuf rosbif rôti', 'boeuf à bourguignon cuit', 'steak haché 15% m.g. cuit', 'boecuf boulettes cuites', 'porc côtelette grillée', 'porc filet rôti maigre cuit', 'porc rôti cuit', 'porc travers braisé', 'veau escalope cuite', 'veau rôti', 'viande cuite sans précision', 'lapin viande cuite', 'poulet cuisse rôtie viande et peau', 'poulet rôti viande et peau', 'blanc de poulet cuit', 'canard viande rôtie', 'canard magret cuit à la poêle', 'dinde viande rôtie', 'dinde ou autre volaille escalope sautée', 'lieu ou colin noir cuit', 'cabillaud cuit au four', 'cabillaud cuit à la vapeur', "sardine à l'huile en conserve égouttée", 'saumon fumé cru', 'saumon cuit à la vapeur', 'thon au naturel en conserve égoutté', 'poisson pané frit', 'surimi en bâtonnets', 'poisson sans précision', 'autre poisson', 'oeuf dur', 'oeuf à la coque', 'omelette nature cuite au beurre', 'oeuf au plat salé'
4	11	fromages	['camembert fromage à pâte molle à croûte fleurie 45% m.g.', 'fromage à pâte molle à croûte fleurie genre camembert allégé $20-30\%$ m.g.', 'coulommiers fromage à pâte molle à croûte fleurie $40-45\%$ m.g.', 'brie fromage à pâte molle à croûte fleurie 45% m.g.', 'reblochon fromage à pâte pressée non cuite 45% m.g.', 'rouy fromage à pâte molle à croûte lavée 40% m.g.', 'autre fromage à pâte pressée cuite', 'comté fromage à pâte pressée cuite 45% m.g.', 'gruyère fromage à pâte pressée cuite 45% m.g.', 'emmental fromage à pâte pressée cuite 45% m.g.', 'parmesan fromage à pâte pressée cuite 45% m.g.', 'roquefort fromage de brebis à pâte persillée 52% m.g.', 'fromage des pyrénées au lait de brebis', 'raclette fromage à pâte pressée non cuite 45% m.g.', 'tome ou tomme fromage à pâte pressée non cuite 45% m.g.', 'fromage de chèvre affiné au lait pasteurisé ou cru', 'fromage sans précision', 'camembert sans précision', 'caprice des dieux fromage à pâte molle 60% m.g.', 'fromage de chèvre mi-sec', 'vache qui rit fromage fondu 50% m.g.', 'mozzarella', 'boursin fromage frais salé aromatisé 70% m.g.'
5	31, 32	eaux, boissons fraîches sans alcool	'contrex eau plate', 'évian eau plate', 'hépar eau plate', 'vittel eau plate', 'volvic eau plate', 'badoit eau gazeuse', 'eau du robinet', 'eau de source', 'eau minérale ou non faiblement minéralisée', 'eau minérale sans précision', 'jus de citron pressé maison', "jus d'orange à base de concentré pasteurisé", 'soda au cola type coca-cola ou pepsi-cola', 'soda au cola light type coca-cola light ou pepsi-cola light', 'sirop aux extraits de fruits à diluer type menthe ou fraise'
6	38	soupes et bouillons	'soupe aux légumes type fait maison', 'soupe aux légumes préemballée à réchauffer', 'soupe poireau pomme de terre type fait maison', 'soupe à la volaille et aux vermicelles préemballée à réchauffer ou déshydratée r', 'soupe à la tomate préemballée à réchauffer ou déshydratée reconstituée', 'soupe au potiron', 'bouillon de volaille déshydraté reconstitué type bouillon de poule knorr'
7	23	légumes (hors pommes de terre)	'betterave rouge cuite', 'carotte cuite', 'carotte crue', 'oignon cru', 'oignon cuit', 'radis rouge cru', 'céleri-rave cru', 'échalote crue', 'chicorée frisée crue', 'endive crue', 'épinard cuit', 'laitue crue', 'endive cuite', 'mâche crue', 'salade verte sans assaisonnement sans précision', 'autre salade sans assaisonnement', 'avocat frais', 'aubergine cuite', 'courgette cuite', 'poivron vert jaune ou rouge cuit', 'tomate crue', 'maïs doux en conserve', 'tomate cerise crue', 'concombre cru pulpe', 'artichaut cuit', 'asperge cuite', 'poireau cuit', 'asperge en conserve', 'brocoli cuit', 'chou de bruxelles cuit', 'chou rouge cru', 'chou vert cuit', 'chou-fleur cuit', 'champignon tout type en conserve', 'champignon de paris sauté', 'haricot vert cuit', 'petit pois en conserve', 'petit pois cuit', 'haricot vert en conserve égoutté', 'haricot vert surgelé cuit', 'macédoine de légumes en conserve', 'petits pois et carottes en conserve', 'légumes mélangés surgelés', 'légume en purée hors pomme de terre non salé', 'légumes cuits sans précision', 'ratatouille niçoise', 'céleri rémoulade'
8	1	pain et panification sèche	'pain baguette', 'pain grillé maison', 'pain courant français boule à la levure', 'pain de campagne ou bis', 'pain complet ou intégral artisanal', 'pain de mie', 'pain aux céréales artisanal', 'autre pain'
9	33	boissons alcoolisées	'vin blanc 11′, 'vin rouge 10', 'vin rouge 11', 'vin rouge 12', 'vin rosé 11', 'champagne', 'vin sans précision', "bière à 4-5 d'alcool type kronenbourg ou 33 export", 'liqueurs', 'whisky', 'vin doux naturel ou vin muté autre que porto type muscat banyuls ou maury', 'apéritif à base de vin ou vermouth type martini', 'pastis prêt à boire (1+5)', 'kir'
10	34	café	'café noir prêt à boire non sucré', 'café expresso non sucré', 'café soluble reconstitué prêt à boire non sucré'
11	30	sucres et dérivés	'sucre blanc', 'sucre roux', 'sucre blanc ajouté au service', 'confiture ou marmelade tout type', 'miel', 'confiture ajoutée au service'
12	41, 26, 8	entremets, crèmes desserts et laits gélifiés, fruits, pâtisseries et gâteaux	['crème dessert au chocolat industrielle type danette ou petits filous', 'chocolat viennois ou liégeois', 'dessert au lait gélifié aromatisé nappé de caramel type flanby', 'crème dessert rayon frais type danette', 'crème dessert rayon frais ou appertisée sans précision', 'flan pâtissier aux oeufs ou à la parisienne', 'mousse au chocolat industrielle rayon frais', 'riz au lait rayon frais', 'abricot frais', 'ananas frais', 'banane fraîche', 'cerise fraîche', 'citron frais', 'fraise fraîche', 'kiwi frais', 'clémentine ou mandarine fraîche', 'melon frais', 'nectarine non pelée fraîche', 'orange fraîche', 'poire non pelée fraîche', 'pomme non pelée fraîche', 'pamplemousse frais', 'prune reine-claude fraîche', 'pêche non pelée fraîche', 'raisin blanc frais', 'salade de fruits frais', 'gâteau sans précision', 'gâteau au chocolat', 'galette des rois feuilletée fourrée à la frangipane', 'gâteau à la crème', 'éclair', 'tarte ou tartelette aux pommes', 'tarte ou tartelette aux fruits', 'crêpe au froment nature'
13	13	beurre	'beurre doux', 'beurre demi-sel sel maxi 3%', 'beurre allégé ajouté au service', 'beurre allégé ajouté à la cuisson', 'beurre doux non allégé ajouté à la cuisson', 'beurre doux non allégé ajoutée au service', 'beurre salé non allégé ajoutée au service', 'beurre salé non allégé ajouté à la cuisson'
14	14	huile	'huile végétale sans précision', 'huile de colza', "huile d'olive vierge", 'huile de tournesol', 'huile mélangée équilibrée type isio 4', "huile d'arachide ajoutée à la cuisson", 'huile de tournesol ajoutée à la cuisson', 'huile de tournesol ajoutée au service', "huile d'olive ajoutée à la cuisson", "huile d'olive ajoutée au service", 'huile mélangée (type isio 4) ajoutée à la cuisson'
15	15	margarine	'matière grasse allégée 60% m.g.', 'matière grasse allégée 55-60% m.g. riche en oméga 3 et 6', 'margarine au tournesol 60% m.g. ajoutée à la cuisson', 'margarine au tournesol 70% ou 80% m.g. ajoutée à la cuisson', 'margarine ordinaire 70% ou 80% m.g. ajoutée à la cuisson'
16	20	charcuterie	'lardon nature cuit', 'jambon cru', 'jambon sec dd (découenné dégraissé)', 'jambon cuit supérieur', 'jambon cuit', 'jambon cuit dd (découenné dégraissé)', 'andouillette réchauffée', 'boudin noir cuit', 'boudin blanc cuit', 'chipolata cuite', 'merguez de boeuf et de mouton cuite', 'saucisson sec', "saucisson à l'ail", 'saucisse de strasbourg', 'saucisse sans précision', 'saucisse sèche', 'rillettes pur porc', 'rillettes', 'pâté de campagne', 'pâté de foie de porc', 'foie gras en conserve', 'pâté de tête ou fromage de tête', 'autre charcuterie'

height

TABLE 1 – *Composition des 16 catégories du déjeuner/dîner créées par connaissance expert (modèle FiltreCollab)*

3.3.2 Etape 3 : modélisation des catégories/utilisateurs et probabilités de tirage des aliments au sein des catégories

Nous avons défini les probabilités de tirage d'un aliment au sein d'une catégorie à partir des distances dans l'espace appris après entrainement W2V, que les catégories aient été constituées par clustering K-means (petit déjeuner) ou par connaissance expert (déjeuner et diner).

Les métriques de performance de la recommandation d'aliments au sein des catégories définies (scores de log-vraisemblance et MRR "Mean Reciprocal Rank" des aliments du jour caché, moyennés sur l'ensemble des individus) sont présentées aux tableaux 2, 3, 4, 5. Le score de log-vraisemblance est donné pour la valeur d'α qui optimise la log-vraisemblance totale des aliments.

Les résultats sont les suivants :

— **Dans le cas du petit déjeuner** :
Le score de log-vraisemblance totale pour notre modèle est :

- **supérieur aux modèles de recommandation aléatoire par catégorie** (+24%) et de recommandation de l'aliment le plus populaire par catégorie (+54%)

Le MRR est :

- **supérieur aux modèles de recommandation aléatoire par catégorie** (+132%) et de recommandation du plus populaire par catégorie (+63%)
- égal à celui d'un modèle BPR ("Bayesian Personalized Ranking")

— **Dans le cas du déjeuner et du diner** :
Le score de log-vraisemblance totale pour notre modèle est :

- **équivalent à celui d'un modèle de recommandation aléatoire par catégorie** (+4%)
- supérieur au modèle de recommandation du plus populaire (+52%)

Le MRR est :

- **supérieur aux modèles de recommandation aléatoire par catégorie** (+135%) et de recommandation du plus populaire par catégorie (+4600%), ce dernier étant très mauvais à cause de la variété des consommations
- équivalent à celui d'un modèle BPR (−6%)

Il semble pertinent de se comparer essentiellement au modèle aléatoire par catégorie pour voir si l'on a "appris". Contrairement au petit-déjeuner, l'apprentissage par W2V des séquences des déjeuners/dîners n'a permis d'apprendre qu'une information partielle. Nous avons vraisemblablement appris des positions relatives des aliments, i.e. leur ordonnancement au sein d'une catégorie (car le MRR fondé sur le rang est supérieur au modéle aléatoire par catégorie), mais pas des distances "absolues" entre les aliments (car la log-vraisemblance fondée sur les distances est équivalente au modèle aléatoire par catégorie).

Cela nous paraît lié à la variété plus importante des aliments consommés au déjeuner et diner et au nombre de combinaisons d'aliments possibles très supérieur, comparé au petit déjeuner.
Concernant la variété des aliments consommés, le repas (déjeuner ou diner) du jour caché est en moyenne constitué à 49% d'aliments déjà consommés dans le passé et 51% d'aliments nouveaux, tandis que le petit déjeuner caché est en moyenne constitué à 89% d'aliments déjà consommés dans le passé et 11% d'aliments nouveaux.

En raison de cette variété des consommations au déjeuner/diner et de données limitées, nous bénéficions donc beaucoup moins de la connaissance de l'historique de consommation d'un utilisateur et les distances entre aliments ont été "moins bien apprises" entre aliments dans le cas du déjeuner/diner, ce qui explique la moindre performance en termes de recommandation. Des conclusions identiques sont tirées de l'étude de l'impact du paramètre α (le facteur qui contrôle le degré de nouveauté/conservatisme) sur le score de log-vraisemblance (totale, aliments nouveaux et aliments déjà consommés), comme détaillé en Annexe B.

On peut noter que dans les 2 cas (petit déjeuner et déjeuner/diner), la performance en termes de score de MRR de notre modèle est comparable à celle du modèle de recommandation par factorisation matricielle (BPR). Cela nous conforte dans l'idée que la moindre performance observée dans le cas du déjeuner/diner est certainement liée au manque de données plutôt qu'à la méthode utilisée, puisque deux méthodes différentes de filtrage collaboratif (notre modèle et la factorisation matricielle BPR) conduisent à des scores de MRR moyen proches.

Il est à noter que dans tous les modèles considérés nous avons raisonné par catégorie définie à l'étape 2 (MRR moyenné sur les catégories, log-vraisemblance calculée pour chaque aliment par catégorie), ce qui questionne la pertinence des catégories constituées et la possibilité d'optimiser la performance en modifiant les catégories.
Des pistes potentielles d'amélioration de la performance de la recommandation pour le déjeuner et le diner sont précisées à la section 3.3.4.

Modèle testé PETIT DEJEUNER	**Paramètres**	**Log-vraisemblance totale** (1)	**Log-vraisemblance "nouveauté"**(2)	**Log-vraisemblance "conservatisme"**(3)
Modèle filtrage collaboratif Word2vec, clustering et tirage selon loi multinomiale	W2V 10 dimensions, 1000 itérations (bibliothèque gensim), K-means (scikit-learn) k=8, probabilités fonction des distances entre utilisateur et aliments, $\beta = 0.05$, $\alpha = 15$(4)	**-9.99**	**-2.46**	**-7.53**
Modèle aléatoire par catégorie, tirage selon loi uniforme	aliments équiprobables dans une catégorie, $\beta = 0.05$	**-13.11**	**-1.73**	**-11.38**
Modèle "le plus populaire" par catégorie	aliment le plus fréquent par catégorie, $\beta = 0.05$	**-21.76**	**-3.56**	**-18.20**

(1) pour tous les aliments consommés le jour caché
(2) pour les aliments nouveaux consommés le jour caché
(3) pour les aliments du jour caché déjà consommés précédemment
(4) valeur d'α qui maximise la log-vraisemblance totale
Log-vraisemblance moyenne sur l'ensemble des utilisateurs

TABLE 2 – *Performance (log-vraisemblance) du modèle FiltreCollab pour le petit déjeuner vs modèles de référence*

Modèle testé DEJEUNER/DINER	Paramètres	Log-vraisemblance totale	Log-vraisemblance "nouveauté"	Log-vraisemblance "conservatisme"
Modèle filtrage collaboratif Word2vec, catégories constituées selon connaissance expert et tirage selon loi multinomiale	W2V 10 dimensions, 9000 itérations (bibliothèque gensim), connaissance expert k=17, probabilités fonction des distances entre utilisateur et aliments, $\beta = 0.01$, $\alpha = 3$ [(1)]	**-31.58**	**19.43**	**-12.14**
Modèle aléatoire par catégorie, tirage selon loi uniforme	aliments équiprobables dans une catégorie, $\beta = 0.01$	**-33.06**	**-18.79**	**-14.27**
Modèle "le plus populaire" par catégorie	aliment le plus fréquent par catégorie, $\beta = 0.01$	**-69.45**	**-36.49**	**-32.96**

(1) valeur d'α qui maximise la log-vraisemblance totale

TABLE 3 – *Performance (log-vraisemblance) du modèle FiltreCollab pour le déjeuner/diner vs modèles de référence*

Modèle testé PETIT DEJEUNER	Paramètres	MRR ou Rang Moyen Réciproque [(1)]
Modèle filtrage collaboratif Word2vec, clustering et tirage selon loi multinomiale	W2V 10 dimensions, 1000 itérations (bibliothèque gensim), K-means (scikit-learn) k=8, probabilités fonction des distances entre utilisateur et aliments	**0.88**
Modèle aléatoire par catégorie et tirage selon loi uniforme	équiprobabilité des aliments au sein d'une catégorie	**0.38**
Modèle "le plus populaire" par catégorie	aliment le plus fréquent par catégorie	**0.54**
Modèle BPR ou "Bayesian Personalized Ranking" (factorisation matricielle)	Bibliothèque lightFM. Hyperparamètres optimaux : nombre de vecteurs latents : 50, taux d'apprentissage : 1, epochs : 5000.	**0.88**

(1) Moyenne du MRR calculé pour l'ensemble des utilisateurs (et MRR calculé au sein de chaque catégorie)

TABLE 4 – *Performance (Rang Moyen Réciproque) du modèle FiltreCollab pour le petit déjeuner vs modèles de référence*

Modèle testé DEJEUNER/DINER	Paramètres	MRR ou Rang Moyen Réciproque [(1)]
Modèle filtrage collaboratif Word2vec, catégories constituées selon connaissance expert et tirage selon loi multinomiale	W2V 10 dimensions, 9000 itérations, connaissance expert k=17, probabilités fonction des distances entre utilisateur et aliments	**0.47**
Modèle aléatoire par catégorie et tirage selon loi uniforme	équiprobabilité des aliments au sein d'une catégorie	**0.20**
Modèle "le plus populaire" par catégorie	aliment le plus fréquent par catégorie	**0.01**
Modèle BPR ou "Bayesian Personalized Ranking" (factorisation matricielle)	Bibliothèque lightFM. Hyperparamètres optimaux : nombre de vecteurs latents : 50, taux d'apprentissage : 1, epochs : 9000	**0.50**

(1) Moyenne du MRR calculé pour l'ensemble des utilisateurs (et MRR calculé au sein de chaque catégorie)

TABLE 5 – *Performance (Rang Moyen Réciproque) du modèle FiltreCollab pour le déjeuner/diner vs modèles de référence*

3.3.3 Etape 4 : Modalités de recommandation d'un repas à un utilisateur à partir des catégories

Les modalités de tirage des aliments par catégorie définies à la section 3.2.5 puis l'application des règles d'association/exclusion des aliments définies par l'étude des motifs fréquents sont censées assurer une recommandation de repas cohérent.

Même si la cohérence d'un petit déjeuner recommandé n'a pas encore été quantifiée (métrique et protocole d'une étude auprès d'utilisateurs réels à définir), les quelques essais tendent à montrer que les règles d'association/exclusion identifiées sont efficaces pour assurer la recommandation d'un petit déjeuner cohérent.

En revanche, en raison de leur nombre élevé (combinatoire élevée), nous ne sommes pas parvenus

à identifier les règles d'associations et exclusions entre catégories/aliments qu'il faudrait appliquer pour assurer la cohérence d'un déjeuner/diner. Assurer la cohérence d'une recommandation de repas issu de tirages d'aliments au sein des différentes catégories semble ainsi un défi majeur du projet.

3.3.4 Perspectives pour améliorer la performance pour le déjeuner/diner

Nous pouvons identifier des pistes d'amélioration de la performance de la recommandation au déjeuner/diner en intégrant de nouveaux éléments :

- **En conservant le modèle FiltreCollab, avec davantage de données :**
  En augmentant la quantité de données d'entrainement d'une part et l'historique de consommation des utilisateurs d'autre part pour réduire la proportion d'aliments nouveaux, mais cela suppose l'obtention de données de consommation issues de nouvelles enquêtes structurées de manière identique à INCA2 (nomenclature des aliments etc). Cette piste d'intégration de nouvelles données issues de l'enquête INCA3 notamment sera explorée prochainement.
- **En changeant de modèle et en exploitant l'ordre de consommation des aliments (une information non encore utilisée) :**
  Nous disposons de l'ordre de consommation des aliments dans l'enquête INCA2, information non encore exploitée jusqu'à présent. Nous faisons l'hypothèse que nous pouvons valoriser le catactère séquentiel d'un repas et générer des séquences de déjeuners/diners (à partir d'un aliment en entrée) par l'utilisation de réseaux de neurones récurrents appris sur les repas de l'enquête INCA2. Avec l'objectif d'améliorer la représentation des aliments et la cohérence d'une recommandation de repas pour le déjeuner/diner, 2 difficultés majeures identifiées précédemment. Ce nouveau modèle appelé **GenSeqRNN** a été testé et les premiers résultats sont présentés à la section suivante.

4 Modèle GenSeqRNN (axe 1) : recommandation par génération de séquences

4.1 Méthode pour le modèle GenSeqRNN

Dans le cadre de l'axe 1, en tant qu'alternative à notre modèle actuel, nous en avons testé l'apprentissage des séquences de déjeuners/diners avec un modèle RNN (Réseau de neurones récurrents) (Goodfellow, Bengio et Courville 2016) appelé **GenSeqRNN**. Contrairement à la précédente, cette approche ne tient pas compte du profil utilisateur, elle n'est pas personnalisée mais elle exploite le caractère séquentiel d'un repas. Les RNN sont conçus pour prendre en compte les dépendances temporelles dans les données en maintenant une mémoire interne, ce qui les rend efficaces pour des tâches comme la traduction automatique, la génération de texte.

Dans un RNN, la génération de séquence consiste à produire une séquence de symboles discrets (ici des aliments) à partir d'une séquence en entrée (ici des aliments). L'objectif est de décoder à partir de l'état caché une distribution de probabilités multinomiale sur les symboles à engendrer (ici les aliments). Une fonction softmax est utilisée pour obtenir une distribution à partir du décodage de l'état caché.
Une fois le réseau appris, la génération se fait en choisissant le symbole le plus probable dans la distribution multinomiale décodée à chaque pas de temps. Ce symbole est ensuite considéré comme entrée au pas de temps suivant et la génération se poursuit itérativement. La supervision se fait à

chaque pas de temps ([6]).

Dans le projet, après prétraitement de séquences qui sont les repas (déjeuners/diners) de l'étude INCA2, nous entrainons un modèle RNN standard (constitué d'une couche d'"embedding" pour la représentation des aliments, d'une cellule RNN et d'un décodeur) sur un ensemble d'apprentissage. Les paramètres sont optimisés sur un ensemble de validation et la performance est évaluée sur un ensemble de test. La fonction de perte d'entropie croisée est utilisée pour apprendre et guider l'entraînement du modèle. Nous évaluons la performance finale du modèle sur l'ensemble test avec cette métrique et une métrique de classification plus interprétable, l'"accuracy" (taux de "prédiction" correcte du prochain aliment). L'architecture et le principe d'entrainement d'un RNN sont expliqués à l'Annexe D.

Pour implémenter notre modèle **GenSeqRNN**, la bibliothèque Python utilisée est PyTorch. Les principaux hyperparamètres sont le nombre de dimensions des "embeddings" (représentations vectorielles des aliments), le nombre de dimensions de l'état caché, le nombre d'itérations. Nous avons comparé une initialisation aléatoire des représentations vectorielles des aliments et une initialisation avec les vecteurs issus de l'apprentissage avec Word2Vec (W2V).

Nous utilisons ensuite le modèle entrainé et optimisé pour la génération d'une séquence (séquence d'aliments qui constituera une recommandation de repas, déjeuner ou diner) à partir d'un élément donné au départ, soit par tirage selon la distribution multinomiale des aliments, soit en choisissant l'aliment avec la probabilité maximale, à chaque étape. La génération se poursuit jusqu'à ce qu'un mot de fin de séquence soit généré ou que le nombre maximum d'itérations soit atteint ([6]).

Nous définissons des **critères d'évaluation** pour notre modèle **GenSeqRNN**, pour évaluer la performance de recommandation d'un repas. Nous évaluons la performance sur l'ensemble test avec une métrique de classification plus interprétable que la fonction de perte d'entropie croisée : il s'agit du taux de "prédiction" correcte du prochain aliment ("accurary") : nous considérons l'"accuracy" et l'"accuracy" dans le top 3 des aliments prédits.

Par ailleurs, de même qu'à l'axe 1, le critère usuel que nous souhiaterions évaluer est la satisfaction de l'utilisateur suite à la recommandation, mais nous ne pouvons pas l'évaluer à ce stade du projet (cela nécessite une étude avec des utilisateurs réels). Toutefois, une composante majeure de la satisfaction utilisateur nous paraît être la cohérence globale du repas recommandé. Nous considérons donc cette dernière comme critère d'évaluation, et nous souhaitons la quantifier ultérieurement par le biais d'une étude avec des utilisateurs. L'approche est illustrée à la Figure 8.

6. Pour cette partie je me suis appuyée sur le cours de M. Vincent Guigue : `https://github.com/vguigue/tuto_deep/blob/main/notebooks/4_1-rnn-gen.ipynb`

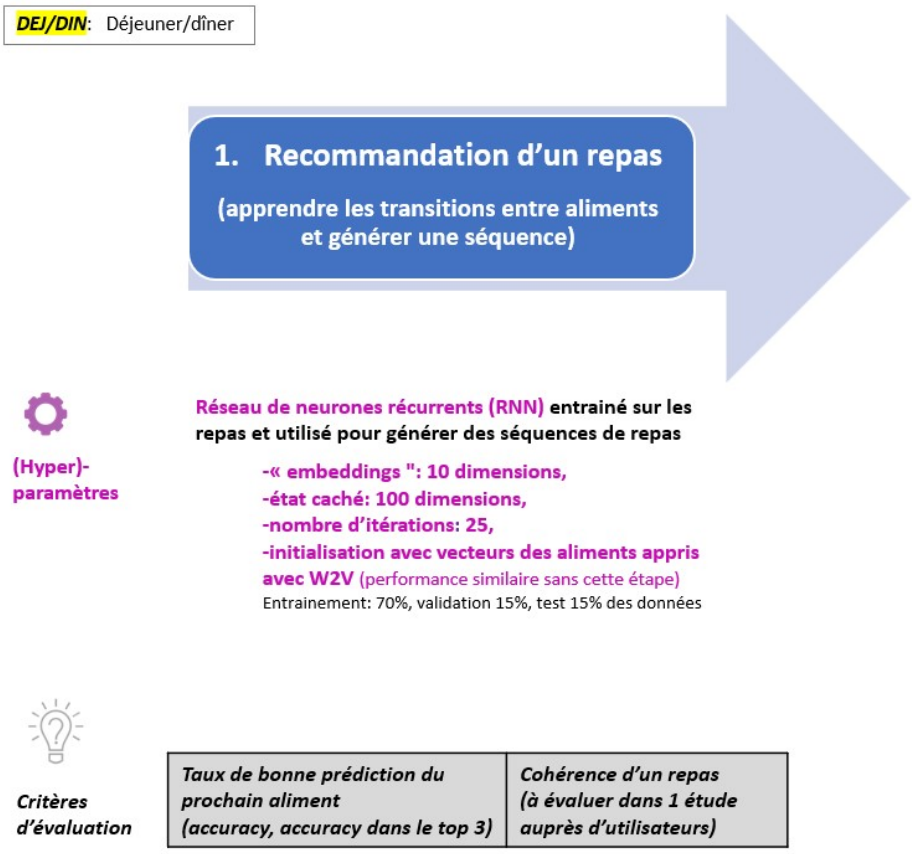

FIGURE 8 – *Paramètres et critères d'évaluation du modèle GenSeqRNN (testé sur le déjeuner/dîner uniquement)*

4.2 Résultats et discussion pour le modèle GenSeqRNN

Le RNN permettant d'apprendre un espace de représentation des aliments, de même que l'algorithme W2V, nous avions fait l'hypothèse que l'initialisation des embeddings des aliments avec les vecteurs issus de l'apprentissage avec l'algorithme Word2Vec pourrait améliorer la performance du modèle, mais cela n'est pas observé : la performance est similaire pour les deux modèles optimisés avec ou sans initialisation W2V, que ce soit en termes de cout de cross entropie ou de métriques d'accuracy pour la prédiction du prochain aliment.

En considérant le modèle initialisé avec les vecteurs W2V et optimisé (10 dimensions d'embedding, 100 dimensions de l'état caché, 25 epochs), les métriques de classification calculées sur l'ensemble test sont les suivantes (en enlevant le mot de fin de séquence qui signale la fin du repas) :

— **"accuracy"** (pourcentage de bonne prédiction du prochain aliment) : **10%**

— **"accuracy top 3"** (pourcentage où l'aliment cible est dans le top 3 des prédictions) : **25%**

Lorsqu'on compare la performance à partir du 4$^{\text{ème}}$ aliment prédit *vs* le premier, le deuxième et le troisième (résultats non présentés ici), une légère tendance à l'amélioration de la performance est observée sur l'"accuracy top3" (ce que l'on attend d'un RNN grâce à la mémorisation de l'état caché), mais elle n'est pas observée sur l'"accuracy".
La performance est meilleure lorsque l'on inclut le mot de fin de séquence mais nous ne l'avons pas retenue car son intérêt dans le domaine applicatif semble moindre (prédiction de la fin du repas).

Ces résultats et en particulier l'"accuracy top 3" sont encourageants pour l'intérêt du modèle

GenSeqRNN pour la recommandation d'un repas déjeuner/dîner. De plus, même si la cohérence des repas générés avec le modèle entrainé n'a pas été quantifiée (étude réelle chez des utilisateurs à conduire), les premiers essais de génération de repas à partir d'un aliment donné en entrée tendent à montrer que des repas cohérents peuvent être obtenus, en appliquant ensuite quelques règles simples d'association/exclusion d'aliments.

Une prochaine étape serait de tester des pistes de personnalisation du modèle **GenSeqRNN** (intégration de l'utilisateur) pour générer des recommandation personnalisées. Ce modèle paraît a priori moins adapté au petit déjeuner en raison de l'absence d'une séquentialité dans les composantes du repas mais cela reste à tester.

4.3 Conclusion sur l'axe 1 du projet

Nous avons testé un premier modèle personnalisé appelé **FiltreCollab** où nous modélisons les utilisateurs au sein d'un espace de représentation des aliments appris avec Word2Vec, où des aliments proches dans l'espace sont substituables. Ce modèle **FiltreCollab** performant sur le petit déjeuner a montré ses limites sur le déjeuner et le diner qui sont caractérisés par une plus grande variété d'aliments (proportion d'aliments nouveaux consommés le jour caché) et un nombre accru de combinaisons d'aliments possibles, avec la même quantité de données.

Il semble que le positionnement des aliments ait été moins bien appris et l'on bénéficie moins de l'historique de l'utilisateur que dans le cas du petit déjeuner. Un défi supplémentaire a été identifié : la difficulté à assurer la cohérence du repas issus du tirage au sein des différentes catégories d'aliments en raison du nombre accru de règles d'association/exclusion des aliments nécessaires.

Disposant de cette information dans l'enquête INCA2, nous avons cherché à valoriser la séquentialité des aliments consommés au sein d'un repas avec un deuxième modèle testé appelé **GenSeqRNN** : l'apprentissage des repas avec un réseau de neurones récurrents où l'on apprend les "transitions" entre aliments et l'on cherche à prédire l'aliment suivant le plus probable. Les résultats du modèle **GenSeqRNN** sonts prometteurs pour la prédiction de séquences de repas : performance mesurée par le taux de bonne prédiction des aliments dans le top 3, cohérence apparente des repas générées (à quantifier dans une étude auprès d'utilisateurs). Des pistes de personnalisation sont à tester pour le modèle **GenSeqRNN**.

Comme illustré à la Figure 9, nous avons donc identifié deux modèles, avec leurs points forts (pour le modèle **FiltreCollab** la personnalisation (BP), qui est un élément clé pour la satisfaction utilisateur, pour le modèle **GenSeqRNN** la prise en compte de la séquentialité (BS) et la cohérence du repas) et leurs limites (défaut de performance sur le déjeuner/diner pour **FiltreCollab**, modèle non personnalisé pour **GenSeqRNN**). Le modèle **GenSeqRNN** paraît a priori moins adapté au petit déjeuner en raison de l'absence d'une séquentialité dans les composantes du repas mais il mérite d'être testé sur ce repas également. L'utilisation de ces deux modèles (en particulier **FiltreCollab**) avec un plus grand nombre de données constitue une autre piste d'amélioration des performances à tester.

Nous avons donc idendifié des méthodes pour répondre à l'objectif initial de recommandation d'un repas basé sur l'apprentissage des consommations des individus. Le défi majeur pour la suite du projet sera de combiner/adapter ces différentes méthodes pour aboutir à un système de recommandation "global" de repas (à l'échelle de la journée et plus).

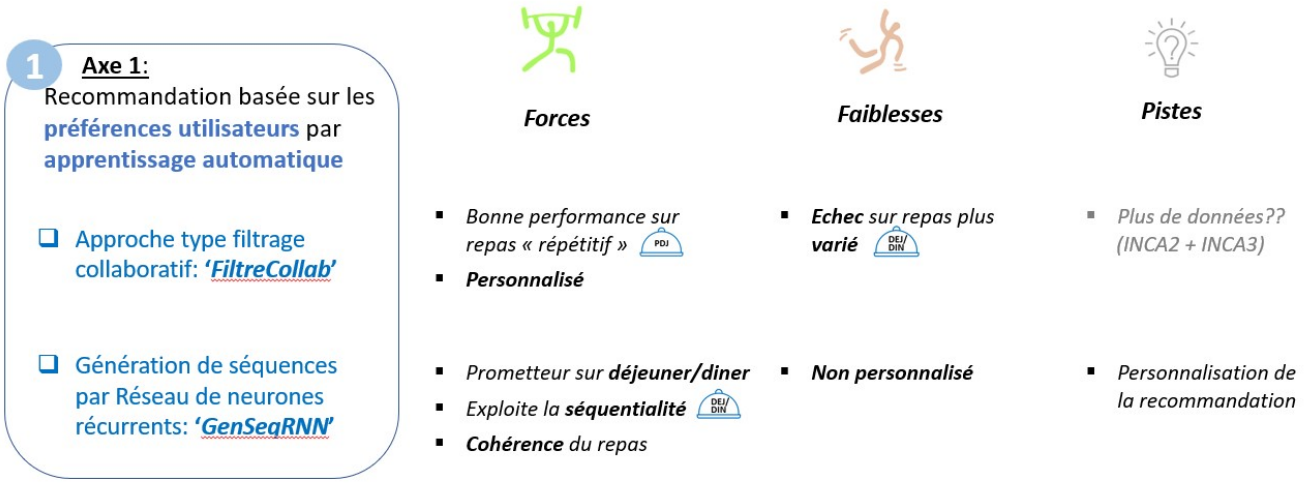

FIGURE 9 – *Axe 1 du projet : forces et faiblesses des 2 modèles testés*

5 Connexion avec l'axe 2 du projet : intégration de règles nutritionnelles à la recommandation

5.1 Rappel de la problématique

L'intérêt de la combinaison des 2 méthodes, i.e. l'intégration des règles nutritionnelles (axe 2) pour filtrer la recommandation issue de l'apprentissage des préférences utilisateurs (axe1), a été détaillé à la section 2.3.4.
La connexion de l'axe 2 avec l'axe 1 a été testée pour le petit déjeuner exclusivement, sur des recommandations issues du modèle **FiltreCollab** présenté à la section 3 (modèle type filtrage collaboratif Word2Vec).

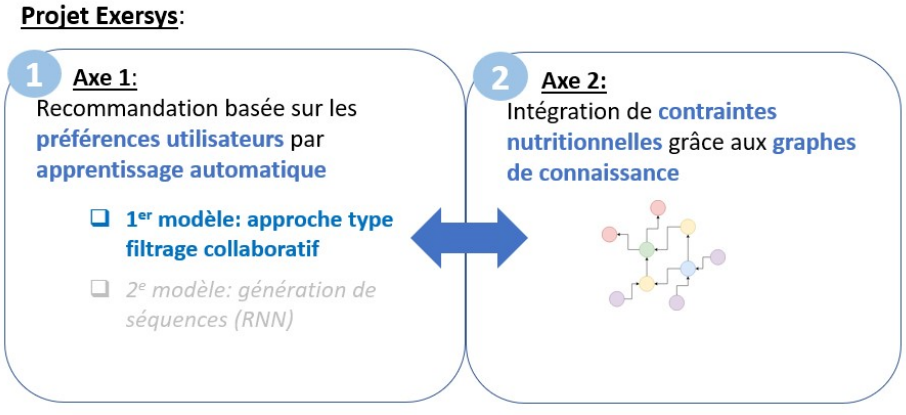

FIGURE 10 – *Connexion testée entre l'axe 2 et le modèle FiltreCollab de l'axe 1, sur le petit-déjeuner exclusivement*

5.2 Méthode : développement de la base de connaissance et des outils pour valider une recommandation (travaux de Ayoub Hammal)

Un stage Master de 2 mois effectué par Ayoub Hammal a porté sur l'axe 2 : il a conduit à la création d'une base de connaissance que l'on appelle **BDNutri** (un ensemble des graphes de connaissances pour modéliser les connaissances sur les aliments et les utilisateurs et le stockage des règles), ainsi que la définition du processus de vérification des règles.

Les graphes sont exprimés sous format RDF (Resource Description Framework, sous forme de triplets), le langage SPARQL (Simple Protocol & RDF Query Language) est utilisé pour interroger le graphe de connaissance et récupérer des informations, et le langage SHACL (Shapes Constraint Language) est utilisé pour valider la conformité des données par rapport à des contraintes définies.

Le rôle de la base de connaissance **BDNutri** est de décrire et relier formellement les informations qui interagissent dans le système de recommandation. Le graphe de connaissance est séparé en unités, avec un grand graphe contenant des informations sur les aliments et un graphe séparé pour chaque consommateur et ses consommations respectives.
La Figure en Annexe E représente le graphe de connaissance global. Les principales classes de ce graphe de connaissances sont celles qui concernent *Food (Aliments)*, *Ingredients* et *Consumers (Consommateurs)*. D'autres informations sont associées aux ingrédients, telles que les allergies qu'ils provoquent et leur origine. Ces informations sont utiles pour définir des contraintes basées sur les allergies et/ou les régimes végétariens, par exemple.

A l'interface des 2 axes, la bibliothèque Python **rdflib** qui permet de manipuler et d'interroger des graphes RDF est utilisée pour faire le lien entre la recommandation de repas émise (axe 1) et l'interrogation du graphe de connaissance (axe 2). L'objectif dans un premier temps est de valider (ou invalider) la conformité d'une recommandation de repas avec la base de connaissances, en vérifiant :

- l'absence d'allergènes et le respect des restrictions alimentaires (végétarisme),
- la conformité avec les règles d'experts (association ou exclusion de certains aliments : par exemple pain si beurre recommandé, pas de poudre soluble si boisson chaude déjà recommandée par ailleurs).

5.3 Résultats et discussion : utilisation de la base de connaissance et des outils développés pour valider une recommandation (mon travail)

Grâce aux délivrables d'Ayoub Hammal j'ai pu :

- émettre une recommandation de repas valide pour un consommateur présentant des allergies ou restrictions alimentaires, en appliquant le principe suivant :
  - en cas de non conformité liée aux allergies, tirage d'un nouvel aliment au sein de la même catégorie que l'aliment interdit qui avait été recommandé,
  - et/ou en cas de non conformité d'associations ou exclusions d'aliments, émission d'une nouvelle recommandation de repas
  - et boucle jusqu'à la conformité avec un nombre maximum de tentatives fixé
- identifier les pourcentages d'aliments problématiques pour chaque catégorie d'aliments, par type d'allergie ou de restriction alimentaire comme illustré à la Figure 11 : cette Figure montre la complexité pour émettre une recommandation pour des profils allergiques au lait ou évitant les produits laitiers

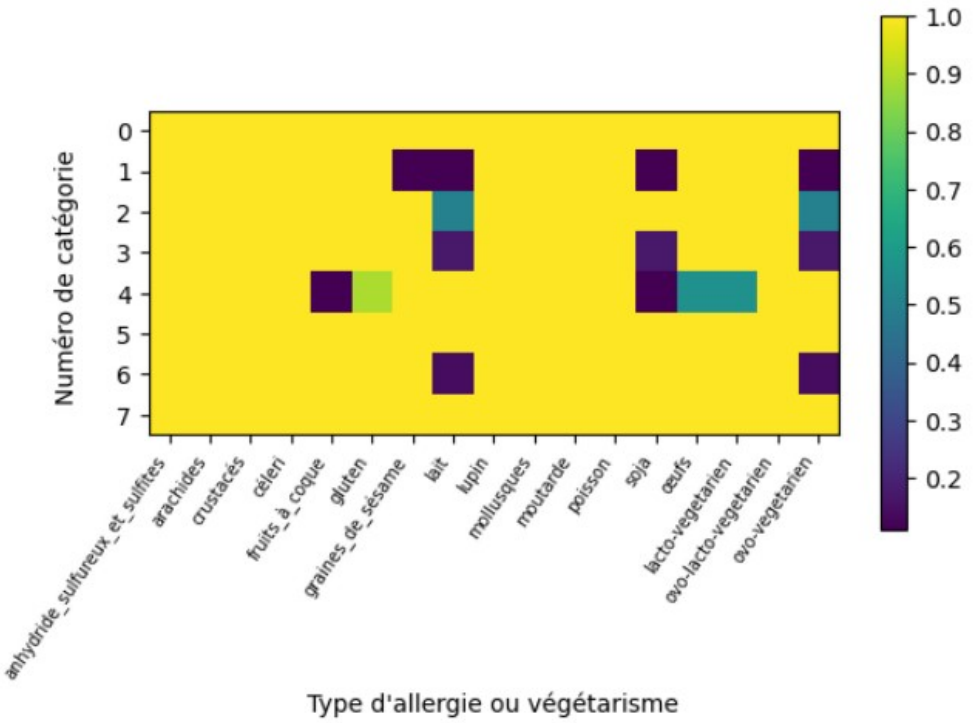

FIGURE 11 – *Pourcentage d'aliments "conformes" identifiés dans chacune des 8 catégories du petit-déjeuner, par type d'allergie ou de végétarisme*

Ces premiers tests illustrent le bon fonctionnement et l'intérêt du filtrage de la recommandation émise à l'axe 1. Des voies d'optimisation peuvent être imaginées à l'avenir :

— optimiser les modalité de retirage des aliments en cas de non conformité (éliminer les aliments interdits de manière définitive pour un utilisateur donné, au lieu de retirer à chaque fois parmi l'ensemble des aliments de la catégorie)

— définir le temps acceptable pour un utilisateur (surtout si le système de recommandation est censé être interactif à l'avenir) pour obtenir une recommandation valide, ce qui permet de fixer le nombre maximum de tentatives acceptables

Ces résultats sont préliminaires et l'ambition du projet à long terme est d'utiliser les graphes de connaissance aussi à d'autres fins, par exemple pour intégrer d'autres élements contextuels du repas (moment, lieu, ...) ou assurer la diversité alimentaire à l'échelle de la semaine.

Conclusion

Pour répondre à l'objectif global du projet qui consistait à développer un système de recommandation de repas basé sur l'apprentisssage automatique des préférences et intégrant des contraintes nutritionnelles grâce aux graphes de connaissances, nous avons abordé à différents besoins et enjeux inhérents à un tel système de recommandation en explorant divers modèles.

Nous avons testé un modèle appelé **FiltreCollab**, basé sur l'apprentissage d'un espace de représentation des aliments avec l'algorithme Word2Vec, la création de catégories d'aliments substituables au sein d'un repas, la modélisation de l'utilisateur au sein des catégories et le tirage d'aliments par catégorie selon une loi multinomiale adaptée qui tient compte des distances apprises. Cette approche a démontré son efficacité pour le petit déjeuner et répond au besoin de personnalisation (**BP**), un objectif crucial dans les systèmes de recommandation.
Cependant, le modèle **FiltreCollab** a montré ses limites lors de l'application au déjeuner/dîner en raison de la plus grande variété d'aliments consommés par les individus et du grand nombre de combinaisons alimentaires possibles, entraînant probablement un apprentissage partiel des distances et une performance réduite.

En exploitant la nature séquentielle (**BS**) du déjeuner/dîner - un aspect absent du modèle initial mais qui façonne intrinsèquement la consommation alimentaire - nous avons testé un second modèle appelé **GenSeqRNN** : un réseau de neurones récurrent entrainé sur les déjeuners/diners INCA2 utilisé pour la génération de séquences. Cette démarche a donné des résultats prometteurs en termes de bonne prédiction du prochain aliment et une meilleure cohérence dans les repas générés. Des pistes pour personnaliser les recommandations ont été identifiées en vue d'explorations futures.

Enfin, en reliant les recommandations de repas émises dans l'axe 1 (modèle **FiltreCollab** appliqué au petit déjeuner) avec l'axe 2 pour intégrer des contraintes nutritionnelles grâce à la base de connaissances **BDNutri** (connexion avec les travaux d'Ayoub Hammal), nous avons démontré la possibilité de prendre en compte des contraintes nutritionnelles (**BN**) telles que les allergies et les préférences végétariennes. La prise en compte de ce type de contraintes est un enjeu clé dans le domaine de la recommandation alimentaire.

Le défi majeur pour les prochaines étapes du projet réside dans la combinaison et/ou l'adaptation de ces méthodes, chacune répondant à des enjeux spécifiques du projet, afin d'aboutir à un système de recommandation de repas complet couvrant l'ensemble d'une journée et au-delà.

Références

Adomavicius, Gediminas et Alexander Tuzhilin (juill. 2005). "Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions". In : *Knowledge and Data Engineering, IEEE Transactions on* 17, p. 734-749. doi : 10.1109/TKDE.2005.99.

Koren, Yehuda, Robert Bell et Chris Volinsky (sept. 2009). "Matrix factorization techniques for recommender systems. IEEE, Computer Journal, 42(8), 30-37". In : *Computer* 42, p. 30-37. doi : 10.1109/MC.2009.263.

Rendle, Steffen et al. (2009). "BPR : Bayesian Personalized Ranking from Implicit Feedback". In : UAI '09. Montreal, Quebec, Canada : AUAI Press, p. 452-461. isbn : 9780974903958.

Mikolov, Tomas, Kai Chen et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv : 1301.3781 [cs.CL].

Mikolov, Tomas, Ilya Sutskever et al. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. arXiv : 1310.4546 [cs.CL].

Delporte, Julien, Stéphane Canu et Alexandros Karatzoglou (2014). "Apprentissage et Factorisation pour la Recommandation". In : *Revue des Nouvelles Technologies de l'Information* Apprentissage Artificiel et Fouille de Données, RNTI-A-6, p. 1-26.

Goldberg, Yoav et Omer Levy (2014). *word2vec Explained : deriving Mikolov et al.'s negative-sampling word-embedding method*. arXiv : 1402.3722 [cs.CL].

Rong, Xin (nov. 2014). "word2vec Parameter Learning Explained". In.

Hidasi, Balázs et al. (2015). "Session-based recommendations with recurrent neural networks". In : *arXiv preprint arXiv :1511.06939*.

Ricci, Francesco, Lior Rokach et Bracha Shapira (jan. 2015). "Recommender Systems : Introduction and Challenges". In : p. 1-34. isbn : 978-1-4899-7636-9. doi : 10.1007/978-1-4899-7637-6_1.

Goodfellow, Ian, Yoshua Bengio et Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Zhang, Fuzheng et al. (2016). "Collaborative Knowledge Base Embedding for Recommender Systems". In : New York, NY, USA : Association for Computing Machinery. isbn : 9781450342322. doi : 10.1145/2939672.2939673. url : https://doi.org/10.1145/2939672.2939673.

Chen, Mingang et Pan Liu (2017). "Performance Evaluation of Recommender Systems". In : *International Journal of Performability Engineering* 13.8, 1246, p. 1246. doi : 10.23940/ijpe.17.08.p7.12461256. url : http://www.ijpe-online.com/EN/abstract/article_3798.shtml.

Gope, Jyotirmoy et Sanjay Jain (mai 2017). "A survey on solving cold start problem in recommender systems". In : p. 133-138. doi : 10.1109/CCAA.2017.8229786.

Trattner, Christoph et David Elsweiler (2017). "Food Recommender Systems : Important Contributions, Challenges and Future Research Directions". In : *CoRR* abs/1711.02760. arXiv : 1711.02760. url : http://arxiv.org/abs/1711.02760.

Caselles-Dupré, Hugo, Florian Lesaint et Jimena Royo-Letelier (2018). *Word2Vec applied to Recommendation : Hyperparameters Matter*. arXiv : 1804.04212 [cs.IR].

Haussmann, Steven et al. (oct. 2019). "FoodKG : A Semantics-Driven Knowledge Graph for Food Recommendation". In : p. 146-162. isbn : 978-3-030-30795-0. doi : 10.1007/978-3-030-30796-7_10.

Lops, Pasquale, Dietmar Jannach et Cataldo Musto (avr. 2019). "Trends in content-based recommendation". In : *User Model User-Adap Inter* 29, p. 239-249.

Wang, Hongwei et al. (2019). "Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation". In : *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA : Association for Computing Machinery, p. 2000-2010. isbn : 9781450366748. doi : 10.1145/3308558.3313411. url : https://doi.org/10.1145/3308558.3313411.

Yera Toledo, Raciel, Ahmad A. Alzahrani et Luis Martínez (2019). "A Food Recommender System Considering Nutritional Information and User Preferences". In : *IEEE Access* 7, p. 96695-96711. doi : 10.1109/ACCESS.2019.2929413.

Guo, Qingyu et al. (2020). "A Survey on Knowledge Graph-Based Recommender Systems". In : *CoRR* abs/2003.00911. arXiv : 2003.00911. url : https://arxiv.org/abs/2003.00911.

Lee, Hea In et al. (2020). "A multi-period product recommender system in online food market based on recurrent neural networks". In : *Sustainability* 12.3, p. 969.

Pecune, Florian, Lucile Callebert et Stacy Marsella (sept. 2020). "A Recommender System for Healthy and Personalized Recipe Recommendations". In.

Hung, PT et K Yamanishi (juill. 2021). "Word2vec Skip-Gram Dimensionality Selection via Sequential Normalized Maximum Likelihood". In : *Entropy* 23, p. 997.

Min, Weiqing et al. (2022). "Applications of knowledge graphs for food science and industry". In : *Patterns* 3.5, p. 100484. issn : 2666-3899. doi : https://doi.org/10.1016/j.patter.2022.100484. url : https://www.sciencedirect.com/science/article/pii/S2666389922000691.

Silva, Vanderlei Carneiro et al. (2022). "Recommender System Based on Collaborative Filtering for Personalized Dietary Advice : A Cross-Sectional Analysis of the ELSA-Brasil Study". In : *International Journal of Environmental Research and Public Health* 19.22. issn : 1660-4601. doi : 10.3390/ijerph192214934. url : https://www.mdpi.com/1660-4601/19/22/14934.

World Health Organization (2023). https://www.who.int/news-room/fact-sheets/detail/healthy-diet.

Santé publique France (s. d.). *50 astuces pour manger mieux et bouger plus*. https://www.mangerbouger.fr/content/show/1501/file/Brochure_50_petites_astuces.pdf.

A Annexe : Présentation de l'algorithme Word2Vec et son architecture

L'algorithme Word2Vec est une méthode efficace pour apprendre des représentations vectorielles de mots ("embeddings") à partir de grandes quantités de données textuelles non structurées (Goldberg et Levy 2014). Ces représentations vectorielles de mots se sont avérées efficaces sur une variété de tâches de traitement du langage naturel.
La base conceptuelle de ce modèle repose sur l'idée que les mots apparaissant dans des contextes similaires ont des significations similaires.

Le modèle Word2vec a été publié en 2013 par une équipe de chercheurs dirigée par T. Mikolov (Mikolov, K. Chen et al. 2013 ; Mikolov, Sutskever et al. 2013. Deux architectures ont été initialement proposées : le modèle de sacs de mots continus (CBOW : continuous bag of words) et le modèle skip-gram. Le CBOW vise à prédire un mot étant donné son contexte, c'est-à-dire étant donné les mots qui en sont proches dans le texte. Un tel contexte est par exemple les x mots à droite et les x mots à gauche du mot à prédire. On parle de taille de fenêtre. Le skip-gram a une architecture symétrique visant à prédire les mots du contexte étant donné un mot en entrée.

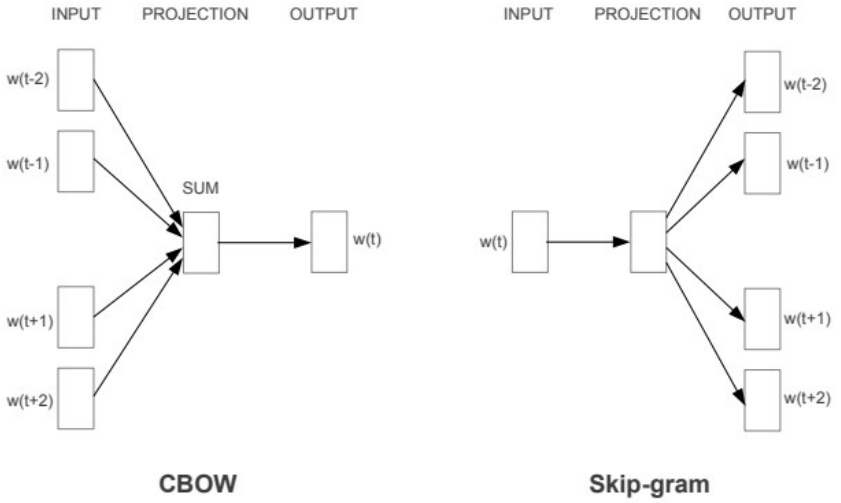

Figure 12 – *Les 2 architectures de Word2Vec (d'après* Goldberg *et* Levy *2014)*

L'objectif d'entrainement du modèle Skip-gram est de trouver des représentations de mots utiles pour prédire les mots environnants dans une phrase ou un document.
Dans ce modèle, étant donné un corpus de mots w et leurs contextes c, on considère les probabilités conditionnelles $p(c|w)$, et étant donné un corpus $Text$, le but est de définir les paramètres θ de $p(c|w;\theta)$ afin de maximiser la probabilité du corpus (Goldberg et Levy 2014) :

$$\arg\max_{\theta} \prod_{w \in \text{Text}} \left(\prod_{c \in C(w)} P(c|w;\theta) \right) \tag{6}$$

Dans cette équation, $C(w)$ est l'ensemble des contextes du mot w.

Alternativement :

$$\arg\max_{\theta} \prod_{(w,c)\in D} P(c|w;\theta) \tag{7}$$

Ici D est l'ensemble de toutes les paires de mots et de contexte que nous extrayons du texte.

Le modèle Word2Vec Skip-gram entraîne un réseau neuronal à 1 couche cachée : une représentation "one-hot-encoding" des mots passe par une « couche de projection » jusqu'à la couche cachée ; ces poids de projection sont ensuite interprétés comme la représentation vectorielle des mots (de dimension égale au nombre de neurones de la couche cachée).

La couche de sortie a autant de neurones que le vocabulaire total. La fonction d'activation de la couche de sortie est généralement softmax, qui traite le problème de la sélection du mot le plus probable comme un problème de classification multiclasse. Etant donné que l'utilisation de softmax représente un coût de calcul élevé, des alternatives peuvent être utilisées : l'approximation « softmax hiérarchique » ou la technique d'échantillonnage négatif [7].

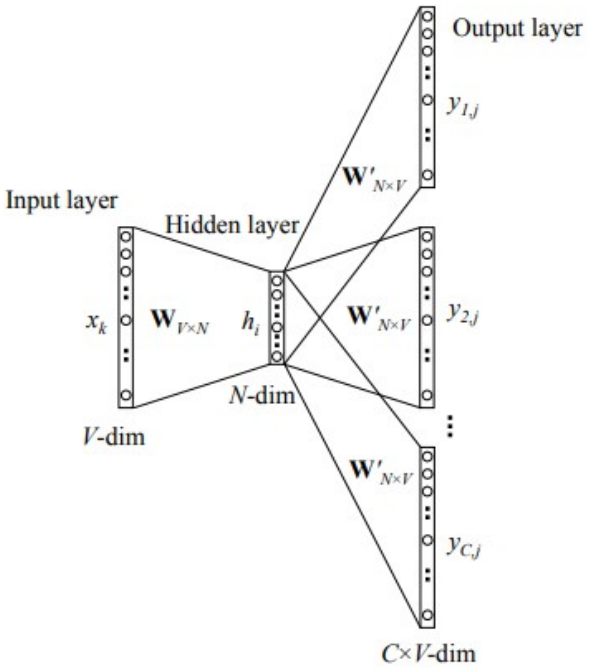

FIGURE 13 – *Le modèle Skip-gram (d'après* RONG *2014)*

7. Avec l'échantillonnage négatif, pour chaque paire mot central-mot contextuel, un certain nombre d'échantillons "négatifs" sont générés. Les échantillons négatifs sont des mots qui n'apparaissent pas dans le contexte du mot central dans les données d'entraînement. L'objectif du modèle devient donc non seulement de prédire le mot contextuel réel, mais aussi de faire la distinction entre le mot contextuel et les échantillons négatifs. Le processus d'entraînement implique désormais de mettre à jour les poids du modèle pour minimiser la probabilité de prédire correctement les échantillons négatifs tout en maximisant la probabilité de prédire correctement le mot contextuel réel.

B Annexe : Impact du paramètre α sur le score de log-vraisemblance des aliments du jour caché

On mesure l'impact de la variation du paramètre α sur la performance de la recommandation de notre modèle en considérant le score de log-vraisemblance moyen sur l'ensemble des individus pour les consommmations du jour caché.
Ce paramètre α est particulièrement intéressant à étudier puisqu'il nous permet de contrôler le degré de nouveauté (faibles valeurs d'α) ou de conservatisme (avec α augmentant) de la recommmandation pour un utilisateur.
On recherche le maximum de la log-vraisemblance du jour caché pour les aliments nouveaux (favoriser la nouveauté), pour les aliments déjà consommés (favoriser le conservatisme), ou pour l'ensemble des aliments (compromis), ces 3 cas correspodant chacun à une valeur d'α optimale :

— respectivement 0, 28 et 15 dans le cas du petit déjeuner cf. Figure 14

— respectivement 0, 12 et 3 dans le cas du déjeuner/diner, cf. Figure 15

On observe la moindre performance de notre modèle dans le cas du déjeuner/diner, où notre modèle n'est pas meilleur que la recommandation aléatoire par catégorie (qui équivaut à $\alpha = 0$) car contrairement au cas du petit déjeuner, on bénéficie peu de l'historique des aliments déjà consommés dans le passé et on a moins bien appris les distances avec la variété alimentaire qui caractérise ces repas.

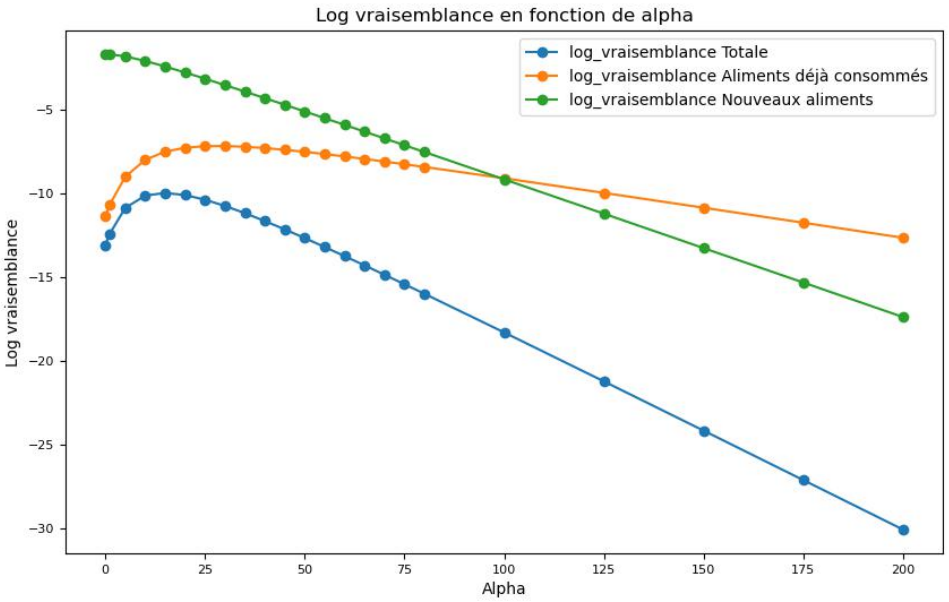

FIGURE 14 – *Log-vraisemblance des aliments du petit déjeuner du jour caché selon notre modèle en fonction de α (moyenne de tous les individus)*

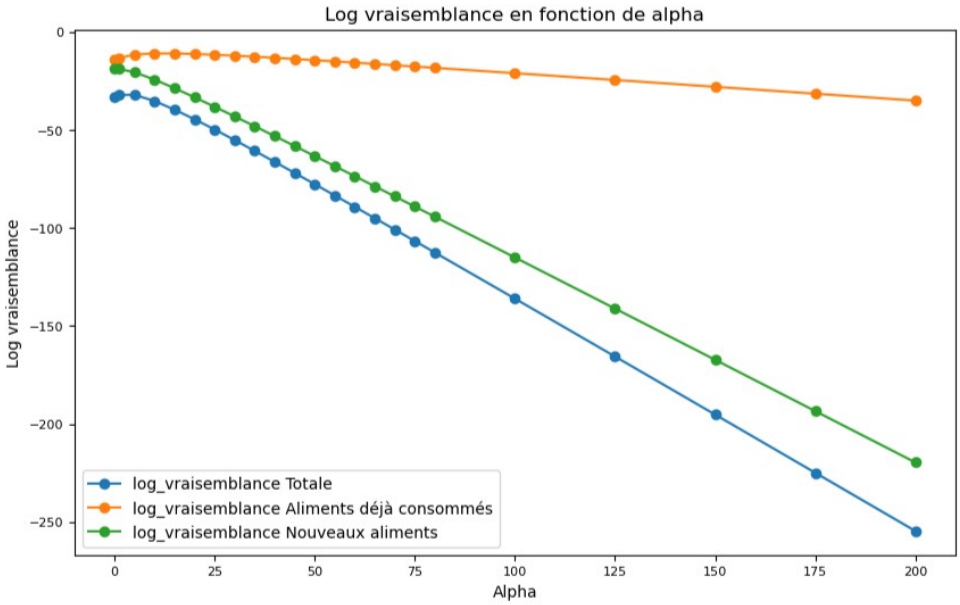

FIGURE 15 – *Log-vraisemblance des aliments du repas (déjeuner ou dîner) du jour caché selon notre modèle en fonction de α (moyenne de tous les individus)*

C Annexe : Processus de recommandation pour le petit déjeuner (modèle 1 de l'axe 1)

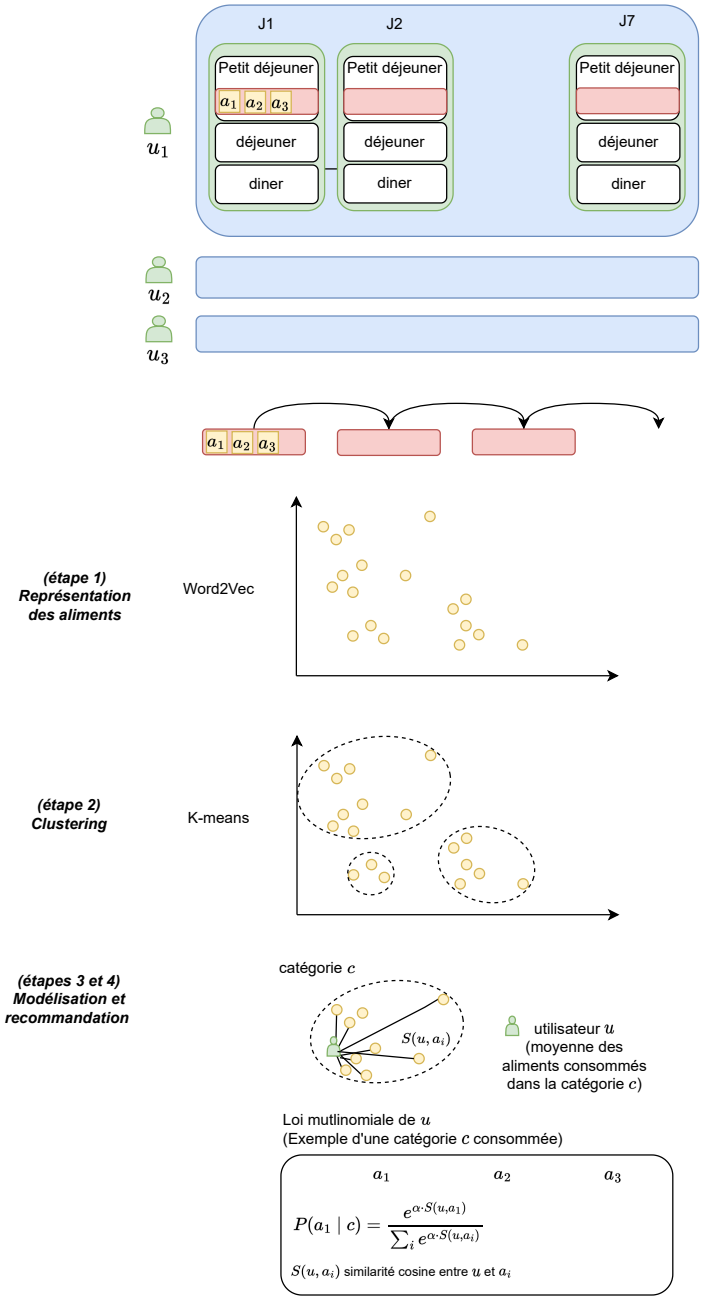

FIGURE 16 – *Processus d'élaboration d'une recommandation défini pour le petit déjeuner*

D Annexe : Présentation des réseaux de neurones récurrents (RNN) : architecture, propagation avant et calcul de la perte d'entraînement

Le réseau de neurones récurrents (RNN) a des entrées vers des connexions cachées paramétrées par une matrice de poids U, des connexions récurrentes « cachées vers cachées » paramétrées par une matrice de poids W et des connexions « cachées vers sortie » paramétrées par une matrice de poids V. La fonction de perte L mesure la distance entre chaque sortie o et la cible d'entraînement y correspondante. la sortie o correspond aux probabilités logarithmiques non normalisées. La perte L calcule en interne $\hat{y} = softmax(o)$ et la compare à la cible y (Goodfellow, Bengio et Courville 2016).

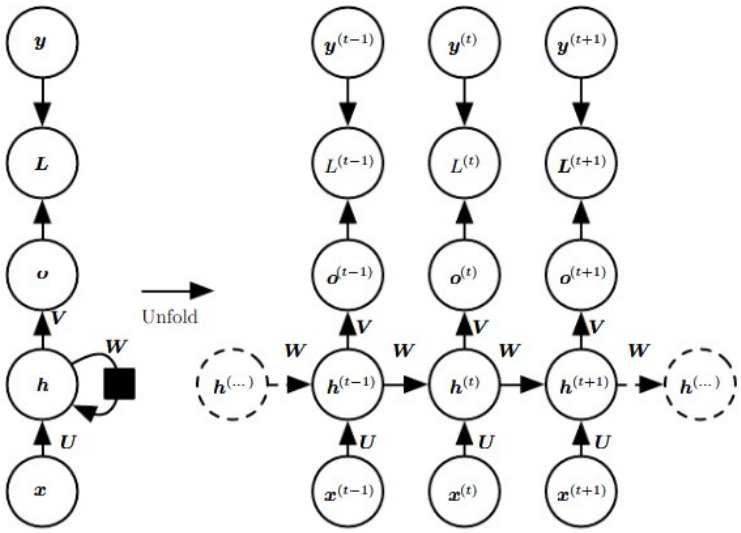

Figure 17 – *Architecture et graphe de calcul de la perte d'entrainement (L) d'un RNN qui mappe une séquence d'entrée de valeurs x à une séquence correspondante de sortie (*Goodfellow, Bengio *et* Courville *2016)*
(À gauche) Le RNN et sa fonction de perte (À droite) La même chose vue comme un graphe déployé dans le temps, où chaque nœud est désormais associé à une instance temporelle particulière

La propagation vers l'avant commence par une spécification de l'état initial $h(0)$. Ensuite, pour chaque pas de temps de $t = 1$ à $t = \tau$, les équations de mise à jour suivantes sont appliquées :

$$a(t) = b + Wh(t-1) + Ux(t) \tag{8}$$
$$h(t) = \tanh(a(t)) \tag{9}$$
$$o(t) = c + Vh(t) \tag{10}$$
$$\hat{y}(t) = \text{softmax}(o(t)) \tag{11}$$

où les paramètres sont les vecteurs de biais b et c ainsi que les matrices de poids U, V et W, respectivement, pour les connexions entrée-cachée, cachée-sortie et cachée-cachée, et h est la fonction d'activation.

E Annexe : Base de connaissance (travaux de Ayoub Hammal)

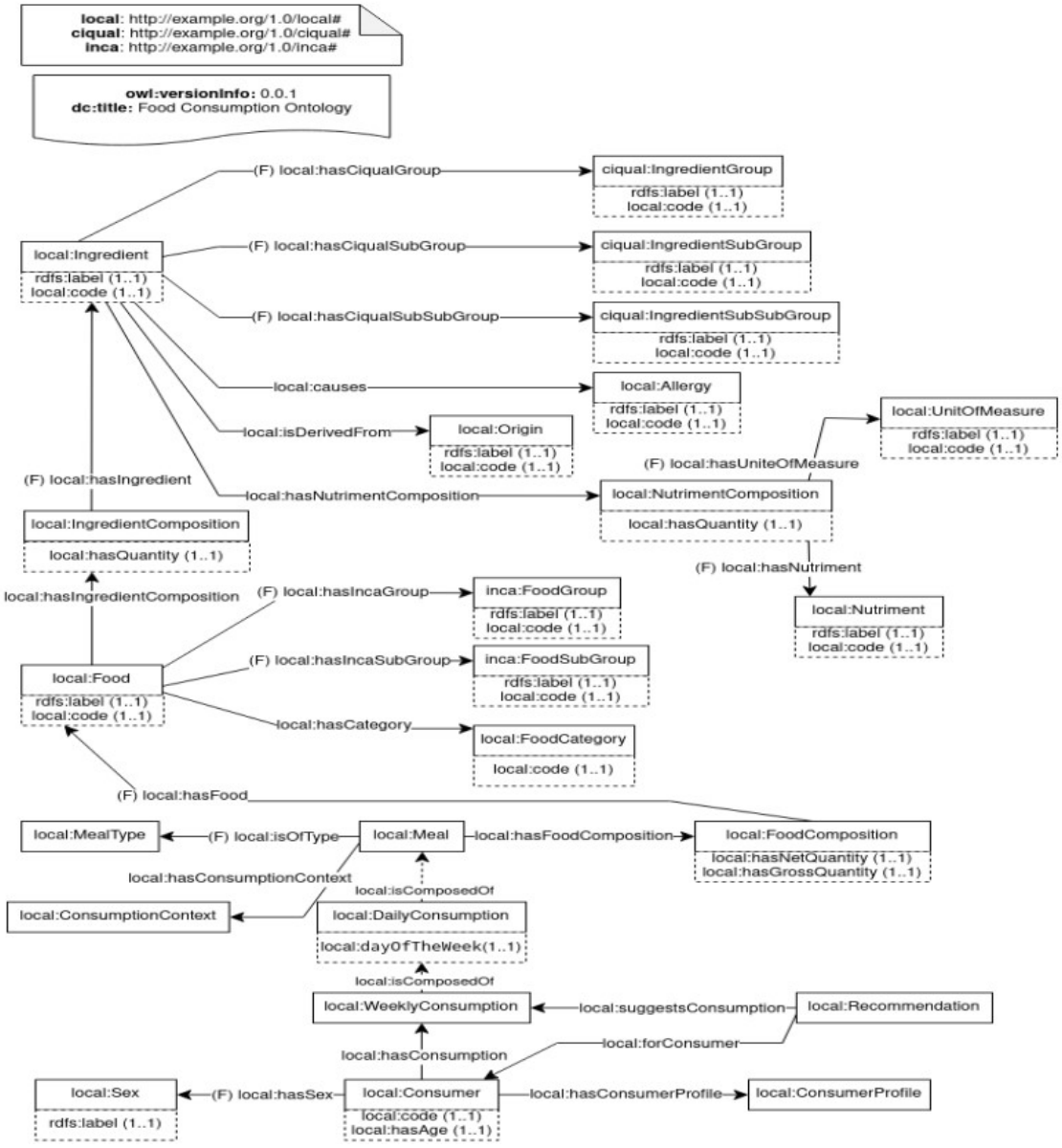

FIGURE 18 – *schéma de la base de connaissance BDNutri créée par Ayoub Hammal (axe 2 du projet)*