Université paris 8

PARIS 8 UNIVERSITE VINCENNES - SAINT-DENIS

Jungle Bike

MÉMOIRE

pour obtenir le grade de Master délivré par

Université Paris 8 Vincennes à Saint-Denis

Mention *Informatique*

Parcours MIASHS Big data et fouille de données

présenté et soutenu publiquement par

Komlan Jean-Marie DANTODJI

le 19 juillet 2022

La recommandation des articles

Encadrant universitaire : Rakia JAZIRI

Tuteur de stage : Alice BATTAREL

Stage effectué à : Jungle Bike
Urban Lab | RIVP Rue René Clair, 75018 Paris

Université Paris 8
Laboratoire d'Informatique Avancée de Saint Denis
EA n° 4383 Saint Denis, France

MASTER

Sommaire

Introduction Génarale

Les systèmes de recommandation aujourd'hui deviennent populaires dans la connaissance des intérêts et préférences des clients qui visitent les sites e-commerce. L'objectif de cette technologie est de comprendre au mieux le client dans ses envies afin de lui proposer encore plus des produits pouvant potentiellement l'intéresser. Les géants de la recommandation aujourd'hui sont Google, Amazon, Netflix,...
L'objectif de cette thèse est donc de proposer une amélioration de la performance des systèmes de recommandations dans le commerce en ligne.

Première partie

Problématique

Sommaire

Chapitre 1

Le contexte de résolution du problème

Sommaire

1.1 Le problème à résoudre

Bien que la recommandation aujourd'hui soit le moyen efficace d'améliorer non seulement la connaissance client mais aussi bien ciblé ses clients, elle pose des problèmes dans sa mise en place. Les grandes difficultés confrontées dans la mise en place des modèles de recommandation :

1. L'information sur l'avis client :
   L'avis du client sur un produit donné peut être explicite soit direct ou implicite. L'avis explicite c'est des formes de likes, votes sur les préférences que le client attribue à un produit après l'avoir testé. Le problème à ce niveau est que peu de clients font des retours d'expérience sur le produits, par conséquent moins d'information pour construire un bon modèle de recommandation.
   D'un autre côté on dispose des informations implicites que le client donne en faisant l'analyse de son comportement. C'est l'exemple du temps passé sur une page, les types d'articles qui ont plus de clics… Les avis implicites sont faciles à collecter en masse car ils ne demandent pas d'effort du côté client. Mais, le problème ici aussi est que ces avis implicites ne renseignent pas effectivement l'intérêt ou non du client à un article. La présence des avis implicite des clients en abondance ont permis de faire beaucoup de recherche dans l'amélioration des modèles de recommandation.

2. Manque d'avis clients en comparaison au nombre d'article à recommander :
   On dispose en majorité peu de données sur le vote et les avis clients en rapport avec le nombre produit présent dans la catalogue.

3. Biais de popularité : non diversité des produits recommandés
   En effet, l'objectif d'un système de recommandation est d'avoir moins de suggestion en haut de la liste recommandée, et induire plus de diversité dans cette liste recommandée. Par contre certains articles nouveaux par faute de popularité auront moins de chance de faire partie de la liste alors qu'ils pourraient potentiellement intéresser le client.

Le problème à résoudre dans ce mémoire sera d'appliquer des modèles de recommandation au secteur du vélo. Les données sont essentiellement des votes que les clients attribuent à chaque pièce de vélo lors de l'achat.

1.2 Présentation des données

1.2.1 Dataset :

Les données à étudier contiennent des informations de chaque article avec le vote, l'avis, ... que le client lui a attribué. Ces données sont issues du scrapping des sites des fournisseurs. Elle contient actuellement 30.912 lignes et 11 colonnes.

	product_id	product_name	brand	user	city	age	activity	level	vote	avis	description
0	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	Br74	Annecy	45-54	VTT - XC	Eclairé	3	Très déçu par le poids réel	Connaissant très bien ce pneu car utilisé en 7...
1	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	StM21	Dijon	45-54	Route - Cyclosportive	Eclairé	5	Très satisfait	Après plus de 15000 km parcourus avec ces pneu...
2	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	boddishiva	barcares	45-54	Route - Cyclosportive	Amateur	2	deçu peut etre un default	j ai acheté ses pneu l an dernier j ai pas par...
3	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	Conti2021	None	35-44	Route - Cyclosportive	Eclairé	4	Une fissure après 1200 km	J'avais fait bcp de bornes avec le GP5000 en 2...
4	0	Pneu Route Continental GP 5000 700 mm Tubetype...	CONTINENTAL	Thibj	Strasbourg, France	25-34	Route - Cyclosportive	Eclairé	1	Mauvaise usure	À peine une dizaine de sortie (courte) et l'on...
...	...	...	...	...	...	...	...	...	...	...	...
30907	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	5	Au top !	[Cet avis a été recueilli en réponse à une off...
30908	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	5	Souple et confortable	[Cet avis a été recueilli en réponse à une off...
30909	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	4	Confort	[Cet avis a été recueilli en réponse à une off...
30910	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	4	chaussure confortable	[Cet avis a été recueilli en réponse à une off...
30911	9221	Chaussures de Trail Femme Asics Gel Sonoma 6 G...	ASICS	None	None	None	None	None	5	Asics gel sonoma trade 6 G TX	[Cet avis a été recueilli en réponse à une off...

30912 rows × 11 columns

FIG. 1.1: Jeu de donnée

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   product_id    30912 non-null  int64
 1   product_name  30909 non-null  object
 2   brand         29285 non-null  object
 3   user          26452 non-null  object
 4   city          23821 non-null  object
 5   age           26277 non-null  object
 6   activity      26277 non-null  object
 7   level         26277 non-null  object
 8   vote          27312 non-null  object
 9   avis          27312 non-null  object
 10  description   27312 non-null  object
dtypes: int64(1), object(10)
memory usage: 2.6+ MB
```

FIG. 1.2: Détail des colonnes

1.2.2 Détail des colonnes de la dataset :

- product name : Nom du produit que le client a acheté
- brand : la marque du produit
- user : représente le nom du client
- city : la ville où habite le client,
- age : la tranche d'âge du client, elle peut nous être à identifier pour chaque produit la tranche d'âge d'individus qui s'y intéressent.
- activity : le type d'activité que le client effectue avec son vélo (Compétition, Voyage, …)
- level : le niveau atteint dans son activité dans la pratique du vélo,
- vote : le score attribué au produit acheté sur le site
- avis : défini le sentiment que porte le client à l'issue de l'achat de l'article,
- description : le commentaire que porte le client sur le produit acheté.

Deuxième partie

État de l'art

Sommaire

Chapitre 2

État de l'art des techniques de recommandation

Sommaire

2.1 Recommandation aléatoire

Bien avant la technologie du machine learning, la recommandation était basée sur des propositions aléatoires des articles au client. Certaines fois les articles les plus populaires sont recommandés. Ce qui pose un problème de personnalisation des produits pour chaque client. Un article peut être populaire mais ne pourra intéresser certains,cela peut devenir contre productif si on se base sur ce type de recommandation. C'est l'exemple d'une recommandation d'articles dépendant du sexe du client.

2.2 Recommandation Personnalisée

Cette méthode consiste à recommander un produit sur la base de ses achats précédents, de ses motifs de recherche. Elle vise à proposer au mieux les produits beaucoup plus successible d'intéresser le client. Mais dans ce cas, si le client n'avait jamais effectué d'achat ou pas assez de commande, il reste difficile de recommander d'autres articles susceptibles de l'intéresser.

2.3 Recommandation Objet (Content-Based filtering CB)

Dans ce type de recommandation, on se base sur les caractéristiques que présente le produit et faire la recommandation sur les caractéristiques des produits que le client a déjà choisi (soit recommander des produits similaires à celui choisi dans le panier).
Pour construire le modèle basé sur la méthode du Content Based, il faut tokenizer les caractéristiques des produits ensuite appliquer les méthodes du TF-IDF. Ceci permet d'augmenter l'importance ou la fréquence des mots clés du produits et de réduire les mots inutiles.
Le problème de cette approche de recommandation Objet, est qu'elle nécessite une connaissance profonde des produits à recommander puisqu'elle est basée sur la description ou le nom du produit.
- Exemple : Considérons deux produits qui ont dans leur description les informations suivantes :
P1: “Pneu souple de qualité”
P2: “Pneu de qualité”
Pour connaître la similarité entre ces deux produits on peut appliquer la méthode de similarité basée sur le calcule du cosinus. Tout d'abord, on calcule la fréquence

des mots clés que contient chaque description, ensuite on projette chaque produit dans un repère n-dimensionnel. Deux vecteurs P1 et P2 de cet espace ainsi constitués sont similaires si et seulement si le cosinus de leur angle est petit.

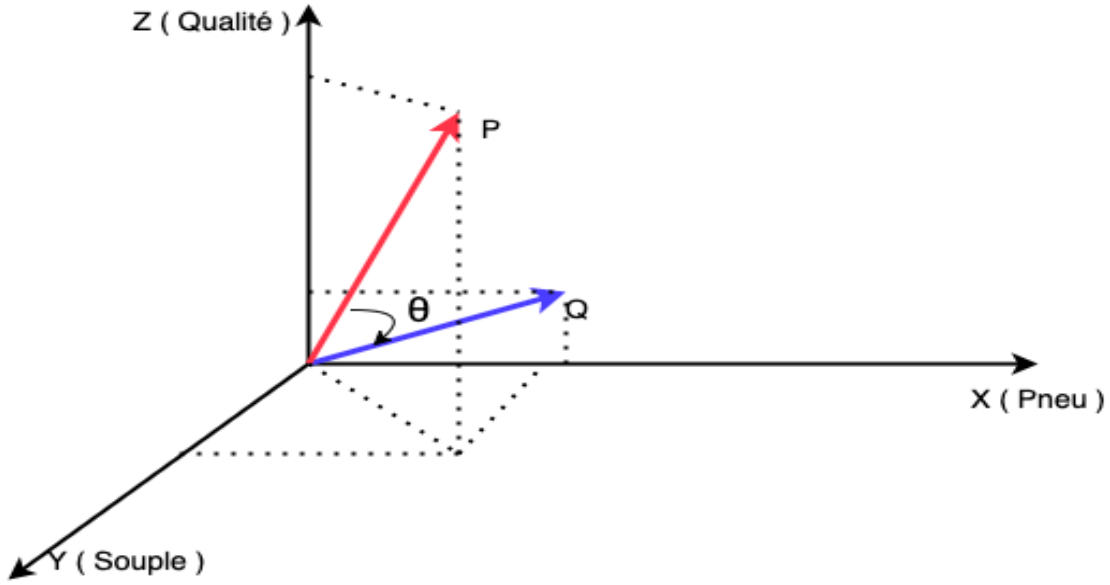

FIG. 2.1: Projection des produits

$$\cos(\theta) = \frac{\mathbf{P} \cdot \mathbf{Q}}{\|\mathbf{P}\|\|\mathbf{Q}\|} = \frac{\sum_{i=1}^{n} P_i Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2}\sqrt{\sum_{i=1}^{n} Q_i^2}}$$

2.4 Recommandation Sociale (Collaborative Filtering CF – Context Aware)

Basée sur le comportement ou le vote des clients, le modèle du Collaborative Filtering utilise les avis clients sur des produits pour les recommander à d'autres utilisateurs.

On distingue plusieurs approches de collaborative filtering :

2.4.1 Memory-based CF

Cette approche se base sur les votes, cliques sur lequel il faut établir une corrélation entre les produits ou entre les utilisateurs afin de recommander un produit quelconque à un utilisateur qui ne l'a jamais vu. Dans ce cas plus précis, les produits recommandés sont ceux achetés par les utilisateurs les plus proches.

2.4.2 La Matrice de Factorisation

Cette méthode vise à factoriser la matrice de base obtenue en considérant le vote de chaque client pour chaque article. Cette factorisation permet de simplifier la matrice de base en deux matrices (Client et Article) dont le produit matriciel est similaire à la matrice de base.

$$M = UxI$$

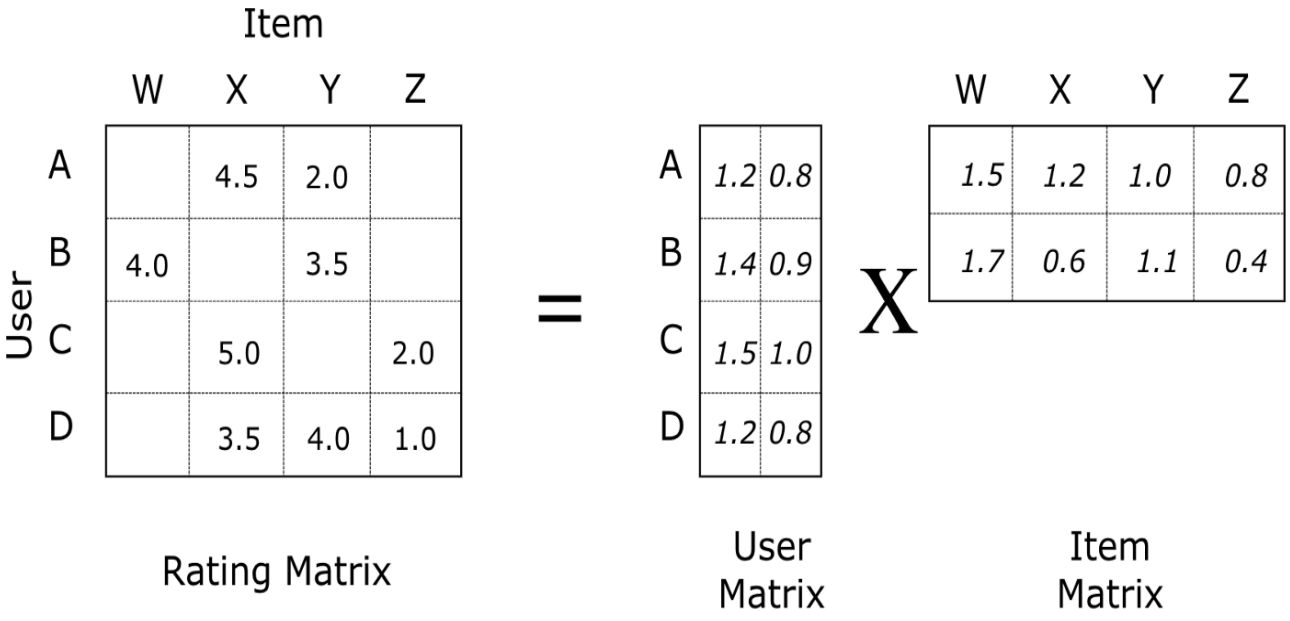

FIG. 2.2: Décomposition de la matrice

3.16	1.92	2.08	1.28
3.63	2.22	2.39	1.48
3.95	2.4	2.6	1.6
3.16	1.92	2.08	1.28

FIG. 2.3: Matrice produit

2.4.3 Neural Collaborative Filtering (NFC)

Il est possible d'appliquer le modèle de réseau de neurone au problème de recommandation. A partir de la matrice des clients et des articles, on envoie en entrée du réseau un encodage de vecteur unitaire du client et de l'article. A l'intérieur le vecteur est connecté à plusieurs couches comme par exemple le perceptron multicouche.

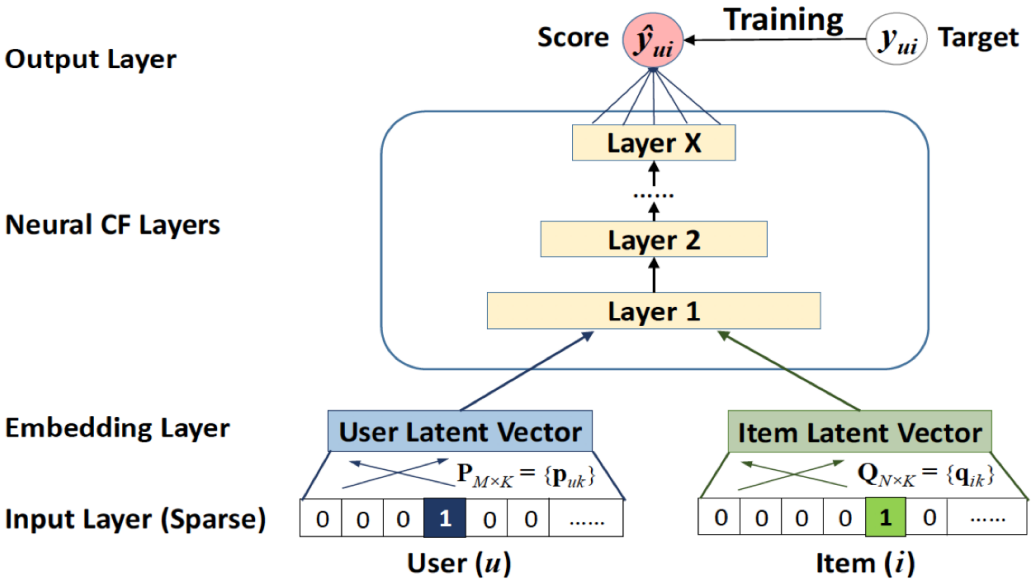

FIG. 2.4: Neural Colaborative Filtering, https ://towardsdatascience.com/neural-collaborative-filtering-96cef1009401

Cette méthode généralise la méthode de factorisation de matrice. Premièrement, en remplaçant la couche interne avec une unique couche de multiplication, on se retrouve avec le schéma ci-dessous.

Ensuite on initialise le poids de la couche de sortie à une matrice J dont toutes les valeurs sont égales à 1 et une fonction d'activation linéaire L.

$$L(x) = x$$

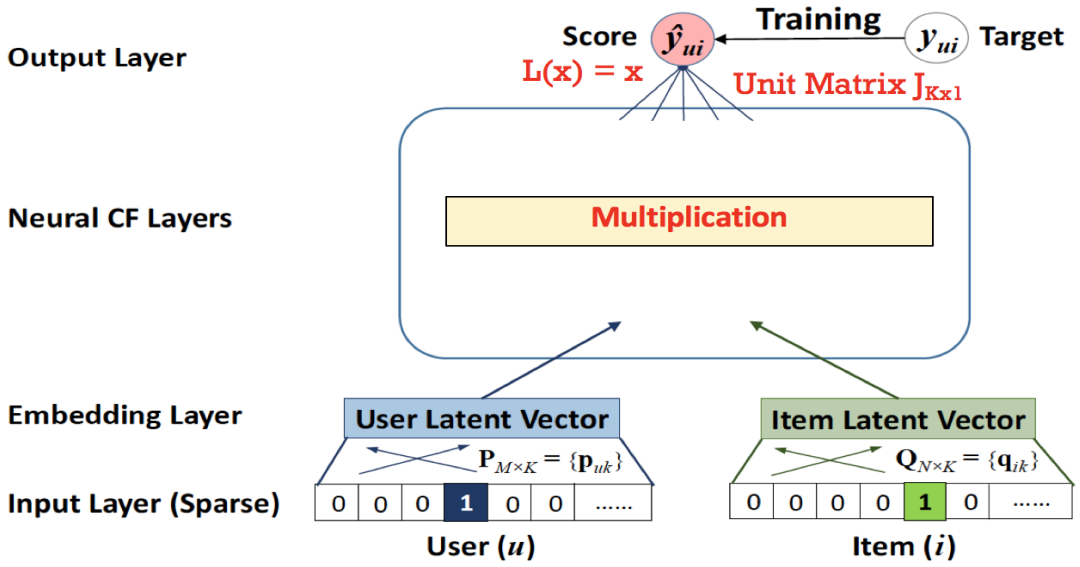

FIG. 2.5: Généralisation du NFC, https ://towardsdatascience.com/neural-collaborative-filtering-96cef1009401

$$\hat{y}_{ui} = L(p_u \odot q_i \times J_{Kx1})$$

$$\hat{y}_{ui} = L(p_u{}^T \cdot q_i)$$

$$\hat{y}_{ui} = p_u{}^T \cdot q_i$$

FIG. 2.6:

Ce qui revient exactement à une décomposition en un produit de deux matrices. On conclut que la méthode de matrice factorisation est un cas particulier du Neuron Collaborative Filtering.

Troisième partie

Conclusion

Sommaire

Conclusion

Le système de recommandation est un moyen efficace pour améliorer le rendement des entreprises et est devenu essentiel dans le secteur du commerce en ligne. Depuis longtemps, la recommandation a connu plusieurs transformations, de la recommandation aléatoire au personnalisée puis basée sur le contenu des produits avec leurs similarités. Aujourd'hui, il en sort plusieurs d'autres approches telle que l'application des modèles de Deep Learning. On a essayé de parcourir toutes ces méthodes qui existent et de comprendre leur fonctionnement. On a compris que la méthode basée sur le Deep Learning est plus efficace car elle permet de généraliser la méthode du Matrice de factorisation. Plus tard, il serait possible de combiner la méthode du Content Based et le Collaborative Filtering dans le but de cibler au mieux les préférences des clients.

Bibliographie

[1] Sumit Sidana. Recommendation systems for online advertising. Computers and Society [cs.CY]. Université Grenoble Alpes, 2018. English. ffNNT : 2018GREAM061ff. fftel-02060436ff

[2] D Gunawan et al. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. 2018 J. Phys. : Conf. Ser. 978 012120

[3] Chakrabarti S, van den Berg M, Dom B 1999 Focused crawling : a new approach to topic-specific Web resource discovery Comput. Networks 31 11–16 pp 1623–1640

Table des figures

Liste des tableaux

Table des matières