Graphes de recommandation enrichis par des informations latentes issues de la factorisation matricielle

Léatitia Ntsamo, Armel Jacques Nzekon Nzeko'o, Jean-François Mehaut, Maurice Tchuente

Graphes de recommandation enrichis par des informations latentes issues de la factorisation matricielle

Léatitia NTSAMO[*] [1,2], **Armel NZEKON**[1,3], **Jean-François MEHAUT**[4], **Maurice TCHUENTE**[1,2,3]

[1]Université de Yaoundé I, Département d'Informatique, BP 812, Yaoundé, Cameroun
[2]Fondation pour la Recherche, l'Ingénierie et l'Innovation, FR2I, BP 14306, Yaoundé, Cameroun
[3]Sorbonne Université, IRD, UMI 209 UMMISCO, F-93143, Bondy, France
[4]Université de Grenoble Alpes, UMR 5217, Laboratoire d'Informatique de Grenoble, France

*E-mail : audrey.ntsamo@facsciences-uy1.cm

Résumé

Les systèmes de recommandation top-N sont importants car ils influencent quotidiennement les choix des utilisateurs des plateformes notamment celles de streaming et de commerce électronique. Les graphes de recommandation sont l'une des approches couramment utilisées pour le calcul des recommandations top-N, car ces derniers sont intuitifs et interprétables. Le principe est de construire un graphe à partir des données explicites des actions des utilisateurs sur les produits, puis d'y appliquer un algorithme de marche aléatoire comme le PageRank pour proposer des recommandations. Cependant, les données implicites qui ne sont pas directement perceptibles dans l'historique des actions des utilisateurs à l'exemple des données latentes résultantes de la factorisation matricielle ne sont pas exploitées dans les graphes de recommandation. En effet, à l'issue d'une factorisation matricielle les utilisateurs et les produits sont représentés dans un nouvel espace défini par les facteurs latents, dans lequel un utilisateur et un item sont proches si l'item est du goût de l'utilisateur. Dans cet article, nous proposons d'exploiter les informations latentes résultantes de la factorisation matricielle pour enrichir les graphes de recommandation. Des expérimentations sont effectuées sur six jeux de données, en utilisant trois métriques d'évaluation des recommandations top-N : Précision, MAP et Hit-ratio. Nos résultats montrent que les graphes enrichis par les informations latentes sont meilleurs dans 97% des cas comparé au graphe initial. L'amélioration des performances peut aller jusqu'à 33% au moins, selon le jeu de données.

Mots-Clés

Graphe de recommandation ; Factorisation matricielle ; PageRank ; Facteurs latents

I INTRODUCTION

Les systèmes de recommandation personnalisés sont de plus en plus pratiques dans les services en ligne tels que le commerce électronique et le streaming. Le but est de proposer à chaque utilisateur de la plateforme concernée, les produits les plus en adéquation avec leurs goûts et préférences. A cet effet, les systèmes de recommandation top-N proposent à chacun d'eux les N produits les plus susceptibles de les intéresser [2]. Par exemple, plus de 80% de l'activité des téléspectateurs de Netflix est motivée par des recommandations personnalisées [1].

1. https ://www.lighthouselabs.ca/fr/blog/how-netflix-uses-data-to-optimize-their-product

Les graphes de recommandation [4] font partie des techniques de recommandation top-N couramment utilisées car ils sont intuitifs et interprétables, l'intérêt pour ces derniers est croissant grâce aux progrès technologiques réalisés sur les tailles des mémoires des ordinateurs. Le principe dans cette approche est de construire un graphe qui représente les actions explicites des utilisateurs sur les produits, puis d'appliquer un algorithme de marche aléatoire comme le PageRank pour calculer les recommandations top-N [6].

Cependant, les graphes de recommandations exploitent essentiellement les données explicites que représentent l'historique des actions des utilisateurs sur les produits, mais ne considèrent pas les données implicites qui ne sont pas directement perceptibles dans cet historique à l'exemple des données latentes résultantes de la factorisation matricielle.

En effet, à l'issue d'une factorisation matricielle, les utilisateurs et les produits sont représentés dans un nouvel espace défini par les facteurs latents, où un utilisateur et un item sont proches si l'item est du goût de l'utilisateur [5, 10]. La prise en compte de telles informations dans les graphes de recommandation pourraient permettre d'améliorer les propositions faites aux utilisateurs. Ainsi, dans cet article, nous proposons d'enrichir les graphes de recommandation par des informations latentes issues de la factorisation matricielle.

La suite de cet article est structurée comme suit : nous présenterons d'abord les systèmes de recommandation de base de notre cadre de travail dans la section II, ensuite nous présentons des techniques d'extension des graphes de recommandation par l'ajout des informations latentes dans la section III. La section IV est consacrée aux expérimentations et à la présentation des résultats. Nous concluons ce document dans la section V.

II BACKGROUND : GRAPHE ET FACTORISATION MATRICIELLE POUR LE CALCUL DES RECOMMANDATIONS TOP-N

Dans une plateforme de commerce électronique ou de streaming qui contient un grand nombre de produits, les systèmes de recommandation ont pour objectif de faciliter l'accès à chaque utilisateur à des produits que ce dernier n'a pas encore sélectionnés et qui vont probablement lui plaire dans un futur proche. Pour atteindre cet objectif, l'approche la plus populaire est celle du filtrage collaboratif qui suppose que « les utilisateurs qui ont eu les mêmes préférences dans le passé auront les mêmes préférences dans le futur ».

Le filtrage collaboratif est également l'approche la plus étudiée dans la littérature au travers des techniques basées sur la mémoire comme les graphes de recommandation [4, 6], et les techniques basées sur un modèle d'apprentissage automatique comme la factorisation matricielle [10]. Dans cette section, nous présentons les notations et format des données utilisés dans ce travail, puis nous poursuivons par la description des graphes de recommandation et terminons par celle de la factorisation matricielle et son application au calcul des recommandations top-N.

2.1 Notations et format des données

Soit R la matrice de taille $n \times m$ issue de l'historique des actions des utilisateurs sur les produits, où n est le nombre d'utilisateurs et m le nombre de produits. Chaque utilisateur u est relié à un ensemble de produits m_u et chaque produit i est relié à un ensemble d'utilisateurs n_i. L'appréciation d'un utilisateur u pour un produit i est représentée dans la cellule $R_{u,i}$ qui peut contenir une valeur binaire $R_{u,i} \in \{0, 1\}$ ou alors une note réelle, $R_{u,i} \in [0, 5]$ attribuée par u.

2.2 Graphes de recommandation

Les graphes de recommandation sont mis en œuvre en exploitant la matrice de données, qu'elle soit binaire ou qu'elle contiennent les notes attribuées par les utilisateurs. Dans cette approche, on modélise d'abord les relations utilisateur-produits en construisant un graphe qui est très souvent le graphe biparti classique, ensuite on applique un algorithme de marche aléatoire à l'exemple du PageRank pour le calcul des recommandations top-N tel que décrit dans la suite.

2.2.1 Construction du graphe biparti classique

Pour construire le graphe biparti classique, chaque utilisateur est représenté par un nœud u et chaque item est représenté par un nœud i. Lorsque $R_{u,i} \geq s$ alors une arête (u, i) est créée entre les nœuds u et i, dont le poids $w_{u,i} = R_{u,i}$, où s est le seuil à partir duquel on estime que l'utilisateur u a une appréciation positive pour l'item i. Dans le cadre de nos expérimentations, s est égal au milieu de l'intervalle de notation et donc $s = 0.5$ dans le cas de la matrice binaire et $s = 2.5$ dans le cas de la matrice des notes réelles.

2.2.2 Calcul des recommandations top-N : PageRank

Après avoir construit le graphe biparti classique, on applique un algorithme de marche aléatoire. Dans ce cas, l'hypothèse de recommandation repose sur la transitivité des goûts des utilisateurs et la proximité de l'utilisateur cible par rapport aux produits qui correspondent à ses préférences. Ainsi, l'objectif est de recommander N produits que l'utilisateur cible n'a pas encore consommés et qui sont les plus proches de lui et de son voisinage le plus immédiat.

Le PageRank est un algorithme de marche aléatoire proposé par les cofondateurs de Google dont l'objectif initial était de classer les pages web par ordre d'importance [7]. Plus tard, le PageRank personnalisé, une version adaptée du PageRank a été proposée pour le calcul des recommandations [8, 1]. Ce dernier permet d'identifier les N produits les plus susceptibles d'intéresser l'utilisateur cible, par une marche aléatoire qui a pour point de départ le nœud u associé à l'utilisateur cible. A l'issue de cette marche aléatoire, l'importance de chaque nœud est stockée dans le vecteur PR calculé de manière itérative en appliquant l'équation suivante :

$$PR = \alpha \cdot M \cdot PR + (1 - \alpha) \cdot d \tag{1}$$

Où M est la matrice de transition du graphe construit, $\alpha \in [0, 1]$ est le facteur d'amortissement de la personnalisation et d le vecteur de personnalisation du PageRank. Pour avoir la liste des N produits à proposer à l'utilisateur cible, les scores dans PR des produits que cet utilisateur n'a pas encore sélectionnés sont triés dans l'ordre décroissant et seuls les N produits ayant les plus grands scores sont retenus et recommandés à l'utilisateur cible.

2.3 Factorisation matricielle

Les modèles de factorisation matricielle (MF) tentent de résumer les données contenues dans la matrice initiale par un nombre restreint de valeurs sans perdre d'informations importantes, notamment les corrélations entre les éléments en ligne ou en colonne et les associations fortes entre les éléments en ligne et ceux en colonne. Dans le cadre des systèmes de recommandation, les éléments en ligne sont les utilisateurs et ceux en colonne sont les produits et donc les modèles de factorisation matricielle résument les informations pertinentes de la matrice R d'entrée en caractérisant à la fois les utilisateurs et les produits par des facteurs latents [5, 10].

Le calcul des recommandations top-N avec un modèle de factorisation matricielle se déroule en deux étapes : la première consiste à déterminer la représentation des utilisateurs et des produits dans un espace latent, et la seconde consiste à déduire les recommandations top-N.

2.3.1 Représentation des utilisateurs et des produits dans un espace latent

A cet étape, on calcule les matrices $\overline{U}$ de taille $n \times k$ et $\overline{I}$ de taille $m \times k$ qui sont respectivement l'ensemble caractérisant les utilisateurs et celui caractérisant les produits dans l'espace latent de dimension k fixée de manière à assurer une bonne approximation de R par $\overline{U} \cdot \overline{I}^T$.

Dans la matrice $\overline{I}$, plus la valeur de la coordonnée $\overline{I}_{i,l}$ est grande, plus le facteur latent l associé, caractérise mieux l'item i. Plus cette valeur est petite, moins l'item i est caractérisé par le facteur latent l. Par ailleurs, dans la matrice $\overline{U}$, plus la valeur du coordonnée $\overline{U}_{u,l}$ est grande, plus l'utilisateur u a de l'intérêt pour les produits caractérisés par le facteur latent l. On en déduit que l'intérêt d'un utilisateur pour un item est modélisé par la proximité de cet utilisateur et de l'item concerné dans l'espace latent résultat de la factorisation matricielle.

2.3.2 Calcul des recommandations top-N

Pour recommander les N produits les plus probables de susciter l'intérêt d'un utilisateur cible u, les modèles de factorisation matricielle estiment la préférence de cet utilisateur u pour tous les produits i qu'il n'a pas encore sélectionnés dans le passé en utilisant l'équation suivante :

$$\overline{R}_{u,i} = \sum_{l=1}^{k} \overline{U}_{u,l} \times \overline{I}^T_{l,i} \tag{2}$$

Les valeurs $\overline{R}_{u,i}$ obtenues pour tous les produits concernées sont triées dans l'ordre décroissant et seuls les N produits associés aux plus grandes valeurs sont recommandés à l'utilisateur cible.

III GRAPHES DE RECOMMANDATION ENRICHIS PAR LES INFORMATIONS LATENTES ISSUES DE LA FACTORISATION MATRICIELLE

Les systèmes de recommandation classiques basés sur les graphes ne tiennent pas compte des informations latentes bien que ces dernières permettent d'estimer les préférences des utilisateurs comme dans le cas de la factorisation matricielle. Pour pallier cette limite, nous proposons de paramétrer les graphes avec les informations latentes résultantes de la factorisation matricielle.

A cet effet, deux stratégies utiles pour matérialiser notre idée sont définies. La première consiste à mettre à jour les poids des relations utilisateur-produits en utilisant des estimations produites par la factorisation matricielle. Et la seconde consiste à ajouter des nœuds qui correspondent chacun à un facteur latent et dont le poids de la relation avec chaque utilisateur ou item est proportionnel à la coordonnée de l'utilisateur ou de l'item suivant le facteur latent associé.

3.1 Pondération MF : mise à jour des poids des relations utilisateur-item par des estimations de la factorisation matricielle

Lorsqu'on applique cette stratégie, on s'intéresse uniquement aux arêtes (u, i) du graphe biparti classique. Si $R_{u,i}$ la préférence de l'utilisateur u pour l'item i dans la matrice des données d'entrée est supérieure ou égale au seuil s (à partir duquel on estime que l'utilisateur u apprécie positivement l'item i), alors le poids $w_{u,i}$ de l'arête (u, i) est donné par l'équation :

$$w_{u,i} = (1 - \beta) \cdot R_{u,i} + \beta \cdot \overline{R}_{u,i} \tag{3}$$

où le paramètre $\beta \in [0, 1]$ permet de calibrer l'impact de l'estimation de la préférence de l'utilisateur u pour l'item i par la factorisation matricielle. L'arête (u, i) est maintenue dans le graphe si son poids calculé $w_{u,i} \geq s$, et est supprimée du graphe dans le cas contraire.

Cette stratégie permet ainsi d'exprimer finement les poids des relations utilisateur-produits grâce aux informations latentes de synthèse de la factorisation matricielle.

3.2 Nœuds latents : ajout dans le graphe des nœuds associés chacun à un facteur latent

Dans cette stratégie, une nouvelle catégorie de nœuds est créée à savoir celle des nœuds latents. Un nœud latent est un nœud qui représente un facteur latent dans le graphe. On a donc, autant de nœuds latents que de facteurs latents. Ainsi, le graphe biparti classique initial tel que présenté sur la figure 1 devient un graphe triparti où chaque nœud latent l est relié à tous les nœuds utilisateur u et les nœuds item i tel que présenté sur la figure 2.

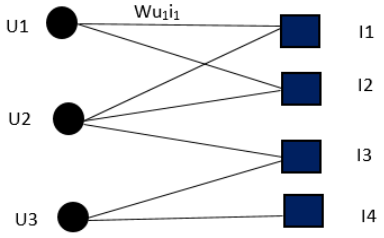

FIGURE 1 – Graphe Biparti

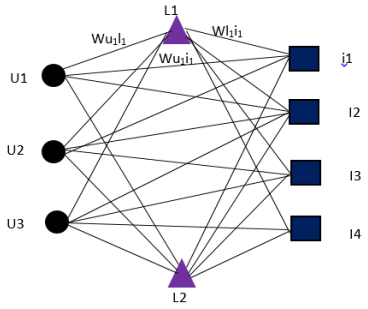

FIGURE 2 – Ajout de la 3ème couche

Les poids des arêtes (u, l) et (i, l) respectivement utilisateur-nœud latent et item-nœud latent sont proportionnels aux coordonnées de l'utilisateur u et de l'item i suivant le facteur latent l. En effet, les coordonnées des utilisateurs et celles des produits dans l'espace latent sont normalisées de sorte que les valeurs maximales soient égales à la note maximale attribuable à un item (1 dans le cas binaire, et 5 dans le cas des notes réelles).

Les poids des arêtes connexes aux nœuds latents sont donnés par les équations suivantes :

$$w_{u,l} = note\_max \cdot \gamma \cdot \frac{\overline{U}_{u,l}}{max(\overline{U})} \tag{4}$$

$$w_{i,l} = note\_max \cdot \gamma \cdot \frac{\overline{I}_{i,l}}{max(\overline{I})} \tag{5}$$

où le paramètre $\gamma \in [0, 1]$ permet de calibrer l'impact de la stratégie Nœuds latents. Par ailleurs, les relations entre les nœuds utilisateur et les nœuds item sont conservées telles que décrites pour le graphe biparti classique dans la sous-section 2.2.1.

La stratégies Nœuds latents permet de rapprocher les utilisateurs et les produits de manière structurelle car il existe toujours un chemin de longueur 2 entre tout couple utilisateur-item. Également de manière préférentielle car plus les coordonnées d'un utilisateur et d'un item sont grandes suivant le même facteur latent, plus le poids du chemin les reliant est important.

3.3 Graphes de recommandation avec les stratégies Pondération MF et Nœuds latents

Dans notre cadre de travail, chaque graphe de recommandation BIP est spécifié par 03 composantes : la première indique la nature de la matrice des données en entrée, la seconde renseigne sur l'usage de la stratégie Nœuds latents et la dernière sur celle de la stratégie Pondération MF.

Nature de la matrice d'entrée. Lorsque la matrice des données d'entrée qui contient les préférences des utilisateurs pour les produits est binaire, le mot clé utilisé est BIN, dans le cas où ce sont des notes réelles attribuées par les utilisateurs, le mot clé utilisé est NTR.

Noeuds latents. Lorsque le système de recommandation utilise la stratégie Nœuds latents telle que décrite dans la sous-section 3.2, le mot clé NDL est présent au niveau de la seconde composante. Dans le cas contraire c'est le mot NA mis pour 'non appliqué' qui est présent.

Pondération MF. De même, lorsque le système de recommandation utilise la stratégie Pondération MF telle que décrite dans la sous-section 3.1, le mot clé PMF est présent au niveau de la dernière composante. Dans le cas contraire c'est le mot NA qui est présent.

Par exemple, BIP-BIN-NDL-PMF est le système de recommandation qui repose sur le graphe biparti classique (BIP), appliqué sur une matrice de données binaires (BIN), avec l'ajout des nœuds latents (NDL) et l'usage de la Pondération MF (PMF). La figure 3 présente la liste de toutes les combinaisons réalisables avec le graphe biparti dans ce cadre de travail.

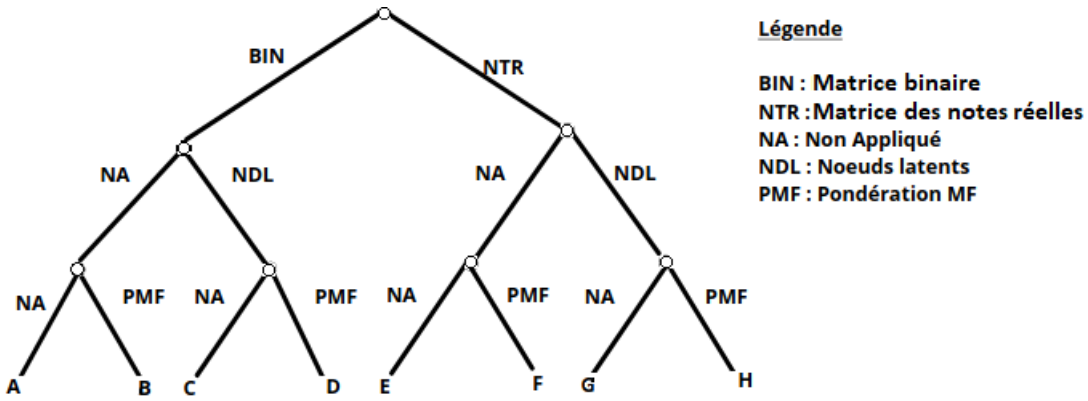

FIGURE 3 – Arbre des combinaisons réalisables avec les stratégies Noeuds latents et Pondération MF.

IV EXPÉRIMENTATIONS

Cette section présente d'abord les jeux de données utilisés, puis le protocole de mise en œuvre et d'évaluation des systèmes de recommandation considérés et est clôturée par la présentation des résultats et commentaires.

4.1 Jeux de données utilisés

Nous utilisons six extraits de jeux de données accessibles publiquement sur internet : movielens-2k [2], movielens-1m [3], ciao, epinions [4], ponpare [5] et retailrocket [6]. Les deux premiers proviennent de MovieLens une plateforme de streaming de films. Les deux suivants, proviennent de Ciao

2. https ://grouplens.org/datasets/hetrec-2011/
3. https ://grouplens.org/datasets/movielens/1m/
4. https://www.cse.msu.edu/~tangjili/trust.html
5. https://www.kaggle.com/c/coupon-purchase-prediction
6. https://www.kaggle.com/retailrocket/ecommerce-dataset

et Epinions, des plateformes où les utilisateurs donnent des avis sur des produits de domaines variés. Le cinquième provient de Ponpare une plateforme de vente en ligne de coupons pour des produits divers et le dernier provient de Retailrocket une plateforme de e-commerce.

La table 1 présente les détails sur les extraits de jeux de données utilisés dans les expérimentations. La première colonne indique la date de début, la seconde la durée en nombre de jours. U et I sont les ensembles des utilisateurs et des produits, et $Nb.(u, i)$ est le nombre de liens distincts utilisateur-produits. Notons que pour avoir les ensembles d'utilisateurs U et d'produits I, un filtre a été appliqué à ceux-ci ; un utilisateur u (resp. un item i) est considéré si le nombre d'occurrences de u (resp. de i) dans l'extrait de jeu de données est supérieur à min_u (resp. min_i).

	Date début	Nb. jours	min_u	min_i	$\|U\|$	$\|I\|$	Nb. (u,i)
Movielens-2k	2008-10-27	70	500	300	236	754	5 497
Movielens-1m	2002-12-16	70	1	1	203	1 590	3 420
Ciao	2010-09-13	210	1	1	299	2 313	2 614
Epinions	2011-02-28	70	1	1	662	5 864	6 203
Ponpare	2012-04-16	70	5	70	364	947	2 683
Retailrocket	2015-07-06	70	100	50	134	4 009	12 718

TABLE 1 – Statistiques sur les jeux de données.

4.2 Protocole de mise en œuvre et d'évaluation des systèmes de recommandation

Pour la mise en œuvre des systèmes de recommandation, les jeux de données sont découpés en deux suivant la dimension temporelle, 80% pour le jeu d'apprentissage et 20% pour le jeu de test. Par exemple, lorsque la durée du jeu de données est 70 jours, les 56 premiers jours contiennent les données du jeu d'apprentissage et les 14 jours restants sont pour le jeu de test.

4.2.1 Paramétrage des systèmes de recommandation

Pour avoir les meilleures valeurs de chaque système de recommandation du cadre de travail, on fait varier les valeurs de tous les paramètres impliqués tel que présenté dans la table 2.

Paramètre	Description	Ensemble de valeurs
k	Nombre de facteurs latents	{5, 10, 20, 50}
α	Calibre la personnalisation du PageRank	{0.25, 0.5, 0.75}
β	Calibre l'impact de la stratégie Pondération MF	{0.25, 0.5, 0.75, 1}
γ	Calibre l'impact de la stratégie Nœuds latents	{0.25, 0.5, 0.75, 1}

TABLE 2 – Ensemble de variation des valeurs des paramètres.

4.2.2 Métriques d'évaluation des recommandation top-N

Pour évaluer les performances des différents systèmes de recommandation top-N, les métriques d'évaluation considérées sont : la 'Précision' qui est la ratio du nombre de bonnes recommandations sur le nombre de recommandations faites, 'MAP' qui est une moyenne des précisions en considérant les rangs des apparitions des bonnes recommandations dans la liste top-N et enfin le 'Hit-Ratio' qui est la proportion des utilisateurs à qui le système a fait au moins une bonne recommandation [9]. Nous utilisons trois valeurs de top-N à savoir : top-5, top-10 et top-20.

4.3 Résultats et commentaires

La table 3 présente tous les résultats des systèmes de recommandation du cadre de travail et suivant tout les top-N et métriques d'évaluation considérés. Pour chaque jeu de données, on a 02 lignes de 09 blocs chacune. Dans la première ligne, le système de recommandation de base est BIP-BIN-NA-NA et dans la seconde c'est BIP-NTR-NA-NA. Dans chaque bloc, le système de base est comparé aux variantes obtenues en utilisant les stratégies Pondération MF et Nœuds latents, mais également au système de recommandation basé sur la factorisation matricielle.

La couleur verte dans une cellule d'un bloc signifie que le système de recommandation en ligne a la meilleure performance suivant la métrique d'évaluation en colonne et est sans ex-æquo. Les autres couleurs bleue, rouge et blanche indiquent l'issue de la comparaison entre les autres systèmes de recommandation et le système de base. Ainsi le bleu indique une amélioration, le rouge une détérioration et le blanc une égalité de performance.

4.3.1 Meilleures performances

Dans la table 3 on a 108 blocs de résultats où le système de recommandation de base est comparé aux variantes résultantes de l'usage des stratégies Pondération MF et Nœuds latents, mais aussi aux résultats de la factorisation matricielle (MF-BIN-NA-NA ou MF-NTR-NA-NA). La table 4 contient 108 cellules associées chacune à un bloc de la table 3 et qui contiennent le système de recommandation qui a la meilleure performance dans le bloc associé. La colonne $am.$ indique le taux d'amélioration ou de détérioration par rapport au système de base du bloc.

La couleur bleue dans la table 4 indique que le meilleur système de recommandation est celui qui utilise uniquement la stratégie Pondération MF (NA-PMF), la couleur verte indique que c'est uniquement la stratégie Nœuds latents (NDL-NA), et la couleur jaune pour la combinaison des deux stratégies (NDL-PMF). Quant à la couleur rouge, elle indique plutôt que c'est la factorisation matricielle qui a produit la meilleure performance (MF). Enfin, la couleur blanche montre qu'aucun système n'a fait mieux que le système de base avec le signe '>' pour dire que le système de base est le meilleur et '==' pour dire qu'il est égalé.

Performances globales. En comparant les performances dans la table 4 des systèmes de recommandation de base à ceux obtenus par application des stratégies Pondération MF et Nœuds latents (NA-PMF, NDL-NA et NDL-PMF), et aux résultats de la factorisation matricielle (MF), on constate que les systèmes de base sont meilleurs à 3% (3/108), la factorisation matricielle est meilleure 14% (15/108), NA-PMF 35% (38/108), NDL-NA 13% (14/108) et NDL-PMF 35% (38/108). On conclut que les graphes de recommandation enrichis par les informations latentes issues de factorisation matricielle sont meilleurs 90 cas sur 108 soit 83% des cas.

Performances en fonction des métriques. Du point de vue métrique d'évaluation, l'usage des stratégie Pondération MF et Nœuds latents conduit à la meilleure performance à 88% (32/36) en Précision, 83% (30/36) en MAP et 77% (28/36) en Hit-Ratio. On peut donc dire que les modèles résultants de notre contribution améliorent mieux le ratio des bonnes recommandations proposées aux utilisateurs que la proportion des utilisateurs satisfaits.

Performances en top-N. En observant les résultats en fonction des top-N, les systèmes de recommandation enrichis sont meilleurs 29/36 (81%) en top-5, 31/36 (86%) en top-10 et 30/36 (83%) en top-20. Ce qui signifie que ces systèmes font mieux que le système de base quelque soit le top pris dans {5, 10, 20} qui sont très utilisés en pratique.

MOVIELENS 2K	Precision						MAP						Hit Ratio					
	P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
BIP-BIN-NA-NA	6.2	-	3.38	-	2.32	-	15.67	-	15.96	-	16.38	-	27.78	-	30.56	-	38.89	-
BIP-BIN-NA-PMF	6.76	9.0	4.08	20.7	2.61	12.5	20.53	31.0	21.1	32.2	21.24	29.7	30.56	10.0	34.72	13.6	41.67	7.1
BIP-BIN-NDL-NA	6.39	3.1	3.61	6.8	2.43	4.7	17.99	14.8	18.39	15.2	18.9	15.4	29.17	5.0	31.94	4.5	40.28	3.6
BIP-BIN-NDL-PMF	6.67	7.6	3.89	15.1	2.57	10.8	20.65	31.8	21.35	33.8	21.6	31.9	30.56	10.0	33.33	9.1	40.28	3.6
MF-BIN-NA-NA	3.06	-51	1.94	-43	1.39	-41	10.28	-35	10.65	-34	10.84	-34	13.89	-50	16.67	-46	20.83	-47
BIP-NTR-NA-NA	6.2	-	3.66	-	2.25	-	16.11	-	16.76	-	17.13	-	27.78	-	33.33	-	38.89	-
BIP-NTR-NA-PMF	7.61	22.7	4.23	15.6	2.75	22.2	21.46	33.2	22.28	32.9	22.71	32.6	34.72	25.0	37.5	12.5	44.44	14.3
BIP-NTR-NDL-NA	6.11	-1.5	3.75	2.5	2.36	4.9	16.3	1.2	17.21	2.7	17.75	3.6	27.78	-	34.72	4.2	41.67	7.1
BIP-NTR-NDL-PMF	6.39	3.1	4.03	10.1	2.5	11.1	18.87	17.1	19.33	15.3	19.72	15.1	29.17	5.0	36.11	8.3	43.06	10.7
MF-NTR-NA-NA	3.06	-51	1.94	-47	1.39	-39	10.28	-37	10.65	-37	10.84	-37	13.89	-50	16.67	-50	20.83	-47

MOVIELENS 1M	Precision						MAP						Hit Ratio					
	P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
BIP-BIN-NA-NA	0.48	-	0.24	-	0.24	-	0.76	-	0.76	-	0.57	-	2.27	-	2.27	-	2.27	-
BIP-BIN-NA-PMF	0.48	-	0.24	-	0.36	50.0	2.27	198	2.27	198	2.42	324	2.27	-	2.27	-	6.82	200
BIP-BIN-NDL-NA	0.91	89.6	0.45	87.5	0.34	41.7	3.03	298	3.03	298	3.03	431	4.55	100	4.55	100	4.55	100
BIP-BIN-NDL-PMF	0.91	89.6	0.68	183	0.45	87.5	3.03	298	3.03	298	3.03	431	4.55	100	6.82	200	9.09	300
MF-BIN-NA-NA	0.45	-6.2	0.45	87.5	0.45	87.5	2.27	198	2.6	242	2.6	356	2.27	-	4.55	100	9.09	300
BIP-NTR-NA-NA	0.48	-	0.24	-	0.48	-	0.76	-	0.76	-	0.76	-	2.27	-	2.27	-	2.27	-
BIP-NTR-NA-PMF	1.43	197	0.71	195	0.6	25.0	3.41	348	3.41	348	2.44	221	6.82	200	6.82	200	9.09	300
BIP-NTR-NDL-NA	0.91	89.6	0.68	183	0.68	41.7	3.41	348	3.41	348	3.35	340	4.55	100	6.82	200	6.82	200
BIP-NTR-NDL-PMF	0.91	89.6	0.68	183	0.68	41.7	4.55	498	4.87	540	3.88	410	4.55	100	6.82	200	9.09	300
MF-NTR-NA-NA	0.45	-6.2	0.45	87.5	0.45	-6.2	2.27	198	2.6	242	2.6	242	2.27	-	4.55	100	9.09	300

CIAO	Precision						MAP						Hit Ratio					
	P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
BIP-BIN-NA-NA	0.11	-	0.06	-	0.11	-	0.11	-	0.11	-	0.23	-	0.56	-	0.56	-	2.22	-
BIP-BIN-NA-PMF	0.11	-	0.17	183	0.14	27.3	0.11	-	0.24	118	0.3	30.4	0.56	-	1.67	198	2.78	25.2
BIP-BIN-NDL-NA	0.22	100	0.17	183	0.17	54.5	0.67	509	0.75	581	0.85	269	1.11	98.2	1.67	198	3.33	50.0
BIP-BIN-NDL-PMF	0.22	100	0.22	266	0.17	54.5	0.74	572	0.88	700	0.88	282	1.11	98.2	2.22	296	3.33	50.0
MF-BIN-NA-NA	0.11	-	0.11	83.3	0.14	27.3	0.56	409	0.62	463	0.62	169	0.56	-	1.11	98.2	2.78	25.2
BIP-NTR-NA-NA	0.22	-	0.22	-	0.17	-	0.42	-	0.57	-	0.66	-	1.11	-	2.22	-	3.33	-
BIP-NTR-NA-PMF	0.67	204	0.34	54.5	0.17	-	2.27	440	2.27	298	2.27	243	3.33	200	3.33	50.0	3.33	-
BIP-NTR-NDL-NA	0.35	59.1	0.22	-	0.17	-	0.67	59.5	0.74	29.8	0.82	24.2	1.67	50.5	2.22	-	3.33	-
BIP-NTR-NDL-PMF	0.67	204	0.38	72.7	0.25	47.1	2.31	450	2.31	305	2.31	250	3.33	200	3.33	50.0	3.33	-
MF-NTR-NA-NA	0.11	-50	0.11	-50	0.14	-18	0.56	33.3	0.62	8.8	0.62	-6.1	0.56	-50	1.11	-50	2.78	-17

EPINIONS	Precision						MAP						Hit Ratio					
	P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
BIP-BIN-NA-NA	0.28	-	0.19	-	0.12	-	0.32	-	0.38	-	0.4	-	1.38	-	1.83	-	2.29	-
BIP-BIN-NA-PMF	0.38	35.7	0.23	21.1	0.16	33.3	0.84	162	0.84	121	0.87	117	1.83	32.6	2.29	25.1	3.21	40.2
BIP-BIN-NDL-NA	0.37	32.1	0.23	21.1	0.14	16.7	0.5	56.2	0.57	50.0	0.6	50.0	1.83	32.6	2.29	25.1	2.75	20.1
BIP-BIN-NDL-PMF	0.37	32.1	0.28	47.4	0.18	50.0	1.07	234	1.15	202	1.18	195	1.83	32.6	2.75	50.3	3.67	60.3
MF-BIN-NA-NA	0.28	-	0.18	-5.3	0.09	-25	0.84	162	0.89	134	0.89	122	1.38	-	1.83	-	1.83	-21
BIP-NTR-NA-NA	0.28	-	0.19	-	0.09	-	0.34	-	0.41	-	0.41	-	1.38	-	1.83	-	1.83	-
BIP-NTR-NA-PMF	0.28	-	0.14	-27	0.14	55.6	0.69	102	0.69	68.3	0.72	75.6	1.38	-	1.38	-25	2.75	50.3
BIP-NTR-NDL-NA	0.28	-	0.18	-5.3	0.11	22.2	0.34	-	0.41	-	0.41	-	1.38	-	1.83	-	2.29	25.1
BIP-NTR-NDL-PMF	0.38	35.7	0.28	47.4	0.14	55.6	0.69	102	0.73	78.0	0.77	87.8	0.92	-34	1.38	-25	1.83	-
MF-NTR-NA-NA	0.28	-	0.18	-5.3	0.09	-	0.84	147	0.89	117	0.89	117	1.38	-	1.83	-	1.83	-

PONPARE	Precision						MAP						Hit Ratio					
	P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
BIP-BIN-NA-NA	0.53	-	0.34	-	0.25	-	1.31	-	1.42	-	1.5	-	2.66	-	3.42	-	4.94	-
BIP-BIN-NA-PMF	0.53	-	0.42	23.5	0.27	8.0	1.39	6.1	1.45	2.1	1.57	4.7	2.66	-	4.18	22.2	5.32	7.7
BIP-BIN-NDL-NA	0.68	28.3	0.46	35.3	0.29	16.0	2.31	76.3	2.53	78.2	2.53	68.7	3.42	28.6	4.56	33.3	5.7	15.4
BIP-BIN-NDL-PMF	0.68	28.3	0.46	35.3	0.29	16.0	2.47	88.5	2.68	88.7	2.68	78.7	3.42	28.6	4.56	33.3	5.7	15.4
MF-BIN-NA-NA	0.68	28.3	0.42	23.5	0.21	-16	2.01	53.4	2.13	50.0	2.13	42.0	3.42	28.6	4.18	22.2	4.18	-16
BIP-NTR-NA-NA	0.53	-	0.34	-	0.25	-	1.31	-	1.42	-	1.5	-	2.66	-	3.42	-	4.94	-
BIP-NTR-NA-PMF	0.53	-	0.34	-	0.27	8.0	1.71	30.5	1.75	23.2	1.88	25.3	2.66	-	3.42	-	5.32	7.7
BIP-NTR-NDL-NA	0.53	-	0.34	-	0.27	8.0	1.55	18.3	1.66	16.9	1.76	17.3	2.66	-	3.42	-	5.32	7.7
BIP-NTR-NDL-PMF	0.61	15.1	0.42	23.5	0.27	8.0	1.71	30.5	1.75	23.2	1.91	27.3	3.04	14.3	4.18	22.2	5.32	7.7
MF-NTR-NA-NA	0.68	28.3	0.42	23.5	0.21	-16	2.01	53.4	2.13	50.0	2.13	42.0	3.42	28.6	4.18	22.2	4.18	-16

RETAILROCKET	Precision						MAP						Hit Ratio					
	P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
BIP-BIN-NA-NA	3.2	-	3.33	-	3.0	-	5.36	-	6.1	-	6.25	-	12.0	-	22.67	-	32.0	-
BIP-BIN-NA-PMF	4.53	41.6	4.13	24.0	3.47	15.7	9.91	84.9	9.55	56.6	9.17	46.7	17.33	44.4	29.33	29.4	33.33	4.2
BIP-BIN-NDL-NA	4.27	33.4	3.87	16.2	3.27	9.0	7.63	42.4	8.32	36.4	8.35	33.6	16.0	33.3	25.33	11.7	32.0	-
BIP-BIN-NDL-PMF	4.53	41.6	4.27	28.2	3.6	20.0	9.19	71.5	10.0	63.9	9.08	45.3	16.0	33.3	29.33	29.4	33.33	4.2
MF-BIN-NA-NA	3.73	16.6	3.6	8.1	2.67	-11	5.86	9.3	6.94	13.8	7.02	12.3	13.33	11.1	25.33	11.7	30.67	-4.2
BIP-NTR-NA-NA	3.2	-	3.33	-	3.0	-	5.36	-	6.1	-	6.25	-	12.0	-	22.67	-	32.0	-
BIP-NTR-NA-PMF	4.27	33.4	4.0	20.1	3.27	9.0	9.2	71.6	8.47	38.9	8.49	35.8	14.67	22.2	26.67	17.6	36.0	12.5
BIP-NTR-NDL-NA	3.2	-	3.47	4.2	3.27	9.0	5.65	5.4	6.56	7.5	6.74	7.8	12.0	-	24.0	5.9	32.0	-
BIP-NTR-NDL-PMF	4.53	41.6	4.27	28.2	3.27	9.0	8.81	64.4	9.31	52.6	8.65	38.4	16.0	33.3	26.67	17.6	33.33	4.2
MF-NTR-NA-NA	3.73	16.6	3.6	8.1	2.67	-11	5.86	9.3	6.94	13.8	7.02	12.3	13.33	11.1	25.33	11.7	30.67	-4.2

TABLE 3 – Récapitulatif des résultats des systèmes de recommandation top-N étudiés.

		Precision						MAP						Hit Ratio					
		P@5	am.	P@10	am.	P@20	am.	M@5	am.	M@10	am.	M@20	am.	H@5	am.	H@10	am.	H@20	am.
MOVIELENS 2K	BIN	NA-PMF	9.0	NA-PMF	20.7	NA-PMF	12.5	NDL-PMF	31.8	NDL-PMF	33.8	NDL-PMF	31.9	NA-PMF	10.0	NA-PMF	13.6	NA-PMF	7.1
	NTR	NA-PMF	22.7	NA-PMF	15.6	NA-PMF	22.2	NA-PMF	33.2	NA-PMF	32.9	NA-PMF	32.6	NA-PMF	25.0	NA-PMF	12.5	NA-PMF	14.3
MOVIELENS 1M	BIN	NDL-NA	89.6	NDL-PMF	183	MF	87.5	NDL-NA	298	NDL-NA	298	NDL-NA	431	NDL-NA	100	NDL-PMF	200	MF	300
	NTR	NA-PMF	197	NA-PMF	195	NDL-NA	41.7	NDL-PMF	498	NDL-PMF	540	NDL-PMF	410	NA-PMF	200	NA-PMF	200	MF	300
CIAO	BIN	NDL-NA	100	NDL-PMF	266	NDL-NA	54.5	NDL-PMF	572	NDL-PMF	700	NDL-PMF	282	NDL-NA	98.2	NDL-PMF	296	NDL-NA	50.0
	NTR	NA-PMF	204	NDL-PMF	72.7	NDL-PMF	47.1	NDL-PMF	450	NDL-PMF	305	NDL-PMF	250	NA-PMF	200	NA-PMF	50.0	-	==
EPINIONS	BIN	NA-PMF	35.7	NDL-PMF	47.4	NDL-PMF	50.0	NDL-PMF	234	NDL-PMF	202	NDL-PMF	195	NA-PMF	32.6	NDL-PMF	50.3	NDL-PMF	60.3
	NTR	NDL-PMF	35.7	NDL-PMF	47.4	NA-PMF	55.6	MF	147	MF	117	MF	117	-	==	-	==	NA-PMF	50.3
PONPARE	BIN	MF	28.3	NDL-NA	35.3	NDL-NA	16.0	NDL-PMF	88.5	NDL-PMF	88.7	NDL-PMF	78.7	MF	28.6	NDL-NA	33.3	NDL-NA	15.4
	NTR	MF	28.3	MF	23.5	NA-PMF	8.0	MF	53.4	MF	50.0	MF	42.0	MF	28.6	MF	22.2	NA-PMF	7.7
RETAILROCKET	BIN	NA-PMF	41.6	NDL-PMF	28.2	NDL-PMF	20.0	NA-PMF	84.9	NDL-PMF	63.9	NA-PMF	46.7	NA-PMF	44.4	NA-PMF	29.4	NA-PMF	4.2
	NTR	NDL-PMF	41.6	NDL-PMF	28.2	NA-PMF	9.0	NA-PMF	71.6	NDL-PMF	52.6	NDL-PMF	38.4	NDL-PMF	33.3	NA-PMF	17.6	NA-PMF	12.5

TABLE 4 – Liste des meilleurs systèmes de recommandation par bloc et améliorations associées.

Performances en fonction de la nature de la matrice. Lorsqu'on considère la nature de la matrice (Binaire-BIN ou Notes réelles-NTR), les systèmes issus de notre contribution sont meilleurs à 93% (50/54) lorsque c'est la matrice binaire, et 74% (40/54) lorsque c'est la matrice des notes attribuées par les utilisateurs. Ce qui veut dire que notre contribution a un impact positif plus important lorsque la matrice des données d'entrée est binaire.

Performances en fonction du jeu de données. Une relecture de la table 4 avec une attention particulière sur les jeux de données permet de constater que les systèmes de recommandation contenant NA-PMF, NDL-NA et NDL-PMF sont meilleurs à 100% (18/18) dans Movielens-2k, 83% (15/18) dans Movielens-1m, 94% (17/18) dans Ciao, 72% (13/18) dans Epinions, 50% (09/18) dans Ponpare et 100% (18/18) dans Retailrocket. On déduit donc qu'à l'exception de Ponpare où notre contribution est meilleure dans 50% des cas, dans tous les autres jeux de données, notre contribution est meilleure au moins 70% des cas et parfois 100%. Ce qui confirme que notre contribution est bonne pour plusieurs domaines.

4.3.2 Conclusion générale sur les meilleures performances

En guise de conclusion, les graphes de recommandation enrichis par des informations latentes issues de la factorisation matricielle sont meilleurs que les graphes de base dans 83% (90/108), et l'amélioration peut aller jusqu'à 33% dans Movielens-2k, 88% dans Ponpare, 84% dans Retailrocket et à plus de 100% dans les jeux de données Movielens-1m, Ciao et Epinions.

Lorsqu'on compare les systèmes de notre contribution à ceux de la factorisation matricielle, on constate qu'ils sont meilleurs dans 84% (91/108). On peut donc conclure que l'ajout des informations latentes dans les graphes de recommandation permet d'avoir des meilleures performances que celle de la factorisation matricielle sources des informations latentes considérées.

V CONCLUSION

L'objectif de ce travail était d'exploiter les informations latentes issues de la factorisation matricielle pour enrichir les graphes de recommandation. A cet effet, deux stratégies sont définies. La première consiste à pondérer les relations utilisateur-produits par des estimations produites par la factorisation matricielle. Et la seconde consiste à ajouter des nœuds qui correspondent chacun à un facteur latent et dont le poids de la relation avec chaque utilisateur ou item est proportionnel à la coordonnée de l'utilisateur ou de l'item suivant le facteur latent associé.

Des expérimentations sont effectuées sur six jeux de données, en utilisant trois métriques d'évaluation des recommandations top-N (Précision, MAP et Hit-ratio) et trois valeurs de top-N (5, 10 et 20). Nos résultats montrent que les performances du PageRank appliqué aux graphes enrichis par les informations latentes sont presque toujours meilleures (105 cas sur 108 - 97%) que celle du PageRank appliqué au graphe initial. L'amélioration des performances peut aller jusqu'à 33% au moins, selon le jeu de données.

Lorsqu'on s'attarde sur les cas où le graphe initial est pondéré par les notes réelles que les utilisateurs ont attribué aux produits (NTR), au moins l'une des variantes des graphes enrichis que nous proposons à un meilleur résultat que la factorisation matricielle dans 40 cas sur 54 (74%). Et lorsqu'on considère un graphe initial binaire (BIN), l'impact positif de notre contribution est encore plus intéressant car les graphes enrichis sont meilleurs 50 cas sur 54 soit 93%.

D'un point de vue pratique, le cadre de travail présenté dans cet article peut être appliqué à une base de données en entrée, en testant toutes les trois variantes de graphes enrichis proposées et en retenant celle qui fournie la meilleure performance pour le critère choisi. Par ailleurs, en guise de perspective, nous proposons de travailler sur la prise en compte des informations latentes dans d'autres systèmes de recommandation comme ceux basés sur les réseaux de neurones ou d'autres algorithmes de classification automatique.

RÉFÉRENCES

[1] Shumeet BALUJA et al. « Video suggestion and discovery for youtube : taking random walks through the view graph ». In : *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, p. 895-904.

[2] Janach DIETMAR et al. *Recommender Systems An introduction*. New York : Cambridge, 2011.

[3] Liu HUAFEND, Jing LINPING et chang MIAOMIAO. « An efficient parallel trust-based recommendation method on multicore ». In : *IEEE Computer Society* 18.10 (2016), p. 9-16.

[4] Zan HUANG, Hsinchun CHEN et Daniel ZENG. « Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering ». In : *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), p. 116-142.

[5] Folasade Olubusola ISINKAYE. « Matrix Factorization in Recommender Systems : Algorithms, Applications, and Peculiar Challenges ». In : *IETE Journal of Research* (2021), p. 1-14.

[6] Armel Jacques NZEKON NZEKO'O, Maurice TCHUENTE et Matthieu LATAPY. « A general graph-based framework for top-N recommendation using content, temporal and trust information ». In : *Journal of Interdisciplinary Methodologies and Issues in Sciences* (2019).

[7] Lawrence PAGE et al. *The PageRank citation ranking : Bringing order to the web.* Rapp. tech. Stanford InfoLab, 1999.

[8] Haveliwala TAHER H. « Topic sensitive pagerank ». In : *ACM* 18.1 (2002), p. 517-526.

[9] Daniel VALCARCE et al. « Assessing ranking metrics in top-N recommendation ». In : *Information Retrieval Journal* 23.4 (2020), p. 411-448.

[10] Koren YEHUDA, Bell ROBERT et Volinsky CHRIS. « Matrix factorization techniques for recommender systems ». In : *IEEE Computer Society* 18.1 (2009), p. 42-49.