Une approche basée sur les motifs graduels pour la recommandation dans un contexte de consommation répétée

Michaël Chirmeni Boujike*, Norbert Tsopze* ****
Jerry Lonlac**, Rosette Nganmeni Njamnou*, Engelbert Mephu Nguifo***, Laure Pauline Fotso*

*Département d'Informatique - Université de Yaoundé 1, Cameroun
michael.chirmeni@facsciences-uy1.cm, norbert.tsopze@facsciences-uy1.cm
fatounjamnou@gmail.com
**IMT Nord Europe, Univ. Lille, Centre for Digital Systems, F-59000 Lille, France
jerry.lonlac@imt-nord-europe.fr
***UCA, CNRS, Mines de Saint-Etienne, LIMOS, 63000 Clermont-Ferrand, France
engelbert.mephu_nguifo@uca.fr
**** IRD, UMMISCO, F-93143, Bondy, Sorbonne University, France

Résumé. Les systèmes de recommandation ont été conçus pour résoudre le problème de surcharge de données. L'objectif est donc de sélectionner parmi un nombre élevé d'items ceux de faible quantité pertinents pour un utilisateur donné. La prise en compte de la nature répétitive et périodique des interactions entre les utilisateurs et les items a permis d'améliorer les performances des systèmes existants. Mais ces systèmes ne prennent pas en compte les données numériques associées à ces interactions. Nous proposons dans cet article une approche de recommandation basée sur les motifs graduels qui permettent de modéliser les covariations entre items. Les résultats expérimentaux obtenus avec l'approche proposée sur le jeu de données utilisé sont encourageants.

1 Introduction

Le développement des technologies a permis à plusieurs entreprises d'orienter une partie de leurs activités sur Internet. Les interactions avec certains clients se font à travers des plateformes conçues pour la cause. Pour le cas du commerce électronique, les entreprises de ce secteur ont des millions de produits dans leurs bases de données, et il n'est pas approprié de proposer un tel nombre d'articles à un client donné au même moment. Ce problème connu sous le nom de surcharge de données est souvent résolu avec les systèmes de recommandation (Schafer et al., 1999). Les chercheurs ont proposé plusieurs approches de recommandation (Parra et Sahebi, 2013) : filtrage par le contenu, filtrage collaboratif, filtrage basé sur les connaissances, ... Ces approches sont généralement conçues pour estimer la probabilité (ou la note) qu'un utilisateur sera intéressé par un item, sans tenir compte de l'aspect répétitif de consommation de la grande quantité des items comme dans le contexte du commerce de détail ou des visites des sites web. Tout récemment, Wang et al. (2019) ont commencé à prendre en compte la consommation répétée dans les recommandations. Il s'agit de tenir compte du fait

qu'un utilisateur, à un moment donné dans le futur, peut toujours être intéressé par un item sur lequel il avait déjà porté son intérêt avant. Dans le cadre de la consommation répétée, les approches existantes (Bhagat et al., 2018; Wang et al., 2019; Lonjarret et al., 2020) étudient plus la variation de la consommation dans le temps. La principale limite de ces approches est la non prise en compte des grandeurs numériques comme les quantités. Pourtant tenir compte de ces informations pourrait améliorer les résultats de la recommandation. Nous proposons une approche nommée GP-RRC (Gradual Patterns for the Recommendation in Repeated Consumption) basée sur les motifs graduels (Di-Jorio et al., 2009; Lonlac et Mephu Nguifo, 2020) pour recommander aux utilisateurs, dans le contexte de consommation répétée. Le choix porté sur les motifs graduels est dû au fait qu'ils permettent de modéliser les covariations (prennent en compte les variations des quantités), et donc permettent de connaître l'intérêt croissant ou décroissant des utilisateurs sur un groupe d'items. Le reste du papier est organisé comme suit : La section 2 fournit un bref aperçu des travaux connexes sur la recommandation dans le contexte de la consommation répétée. La section 3 présente formellement notre modèle GP-RRC. La section 4 présente les résultats expérimentaux sur un jeu de données du monde réel.

2 Travaux de l'état et problème

2.1 Etat de l'art

La question est celle de savoir quel item pourra intéresser l'utilisateur une fois de plus ? Cette problématique a été abordée sous différents angles et dans divers domaines (requêtes de recherche Web répétées (Teevan et al., 2007), écoute de music (Kapoor et al., 2015), E-commerce (Benson et al., 2016) etc...). Certains travaux existants (Anderson et al., 2014; Kapoor et al., 2015; Benson et al., 2016; Chen et al., 2016) se concentrent sur la compréhension du comportement des utilisateurs ; Kapoor et al. (2015) supposent que le comportement de l'utilisateur est guidé par deux états psychologiques latents pour les items : la sensibilisation et l'ennui. Ils sont motivés par l'observation que, dans l'ensemble de données musicales Last.fm, en moyenne chaque utilisateur interagit avec 23% des nouveaux éléments, et 77% des activités sont avec les éléments familiers. Pour atteindre leurs objectifs, ils modélisent explicitement les écarts entre les consommations des utilisateurs de l'item en fonction de ces états latents à l'aide d'un modèle semi-markov caché (HSMM). Anderson et al. (2014) supposent que le comportement des utilisateurs dans le temps est un mélange de comportement répété et de recherche de nouveauté. Basé sur ce principe, ils proposent un modèle qui combine deux facteurs clés : la qualité (qualité de l'item ou son attractivité) et récence (fait que l'utilisateur a des antécédents avec l'item). Selon Chen et al. (2016), ces deux facteurs ne suffisent pas pour modéliser la consommation répétée ; il faut également tirer parti des caractéristiques comportementales. C'est ainsi que Chen et al. (2016) construisent un modèle appelé **TS-PPR** (**T**ime **S**ensitive **P**ersonnalized **P**airwise **R**anking).

D'autres travaux (Bhagat et al., 2018; Wang et al., 2019; Lonjarret et al., 2020) se sont attaqués formellement à la modélisation de la consommation répétée. Bhagat et al. (2018) définissent formellement le problème de recommandation d'achat répété comme suit : compte tenu de l'historique d'achats de produits d'un client (y compris les achats répétés), comment estimer la probabilité que le client répète l'achat d'un produit en fonction du temps écoulé depuis son dernier achat de ce produit. Pour atteindre leurs objectifs, Bhagat et al. (2018) pro-

posent une modification du modèle Poisson Gamma et appellent le nouveau modèle le Poisson Gamma Modifié. Wang et al. (2019) se concentrent sur la construction d'un modèle unifié et holistique qui peut simultanément recommander des articles consommés et de nouveaux articles, où les articles consommés ne sont correctement sélectionnés qu'au bon moment. Pour atteindre leur objectif, Wang et al. (2019) proposent un algorithme nommé SLRC (**S**hort-Term et **L**ife-Time **R**epeat **C**onsumption) basé sur le principe que les consommations de la plupart des articles sont susceptibles de déclencher la prochaine consommation du même article à court terme (effet à court terme) et que certains articles ont tendance à être à nouveau consommés de manière centralisée après un certain temps (effet à vie). Récemment, Lonjarret et al. (2020), ont construit le modèle REBUS (Recommendation Embedding Based on freqUent Sequences) qui utilise des séquences fréquentes pour identifier la partie de l'historique de l'utilisateur la plus pertinente pour la recommandation et utilisent ces séquences pour estimer des chaînes de Markov d'ordres variables et un modèle d'intégration métrique unifié basé sur les préférences de l'utilisateur et la dynamique séquentielle de l'utilisateur. L'utilisation de motifs séquentiels personnalisés permet de sélectionner la partie de l'historique de l'utilisateur récent qui présente le plus d'intérêt pour la recommandation.

2.2 Formalisation du problème

Soit $\delta = (\mathcal{U}, \mathcal{I}, \mathcal{T})$ une base de données de consommation, où $\mathcal{U}$ représente l'ensemble d'utilisateurs, $\mathcal{I}$ l'ensemble d'items et $\mathcal{T}$ l'ensemble de temps. Soit $\mathcal{P}$ l'ensemble de toutes les séquences de consommations dans δ ; chaque consommation est constituée d'un item $i \in \mathcal{I}$ et du temps correspondant $t \in \mathcal{T}$. La séquence de consommation de l'utilisateur $u \in \mathcal{U}$ est donnée par $\mathcal{P}_u = \{(i_1, t_1), (i_2, t_2), ..., (i_{N_u}, t_{N_u})\}$ avec $\mathcal{P}_u \in \mathcal{P}$ et contenant N_u consommations. Posons $\mathcal{P}_u^t$ la séquence de consommation de l'utilisateur u pendant la période allant du temps initial au temps t.

Definition 1 *La consommation (i, t) pour un utilisateur u est dite répétée si et seulement s'il existe une autre consommation $(i, t_0) \in \mathcal{P}_u^t$ avec $t_0 < t$ (Wang et al., 2019).*

Definition 2 *Étant donné un utilisateur u au temps t et son historique d'achat (avec la quantité de chaque item) avant t, le problème est de prédire les items (y compris ceux de son historique d'achat) susceptibles de l'intéresser.*

Chaque quadruplet (u, i, t, x) indique qu'au temps t, l'utilisateur u a consommé ou a acheté x éléments de l'item i. Dans Wang et al. (2019), ce problème est traité, mais les auteurs ne considèrent pas les quantités x. Alors que la prise en compte de ces quantités pourrait améliorer la qualité de la recommandation. L'approche proposée ici prend en compte les quantités à travers les motifs graduels qui modélisent les covariations des grandeurs numériques.

3 Approche proposée

L'objectif principal de cette approche est de recommander à un utilisateur u donné et à un temps t, les items qui pourront l'intéresser (y compris ceux de son historique).

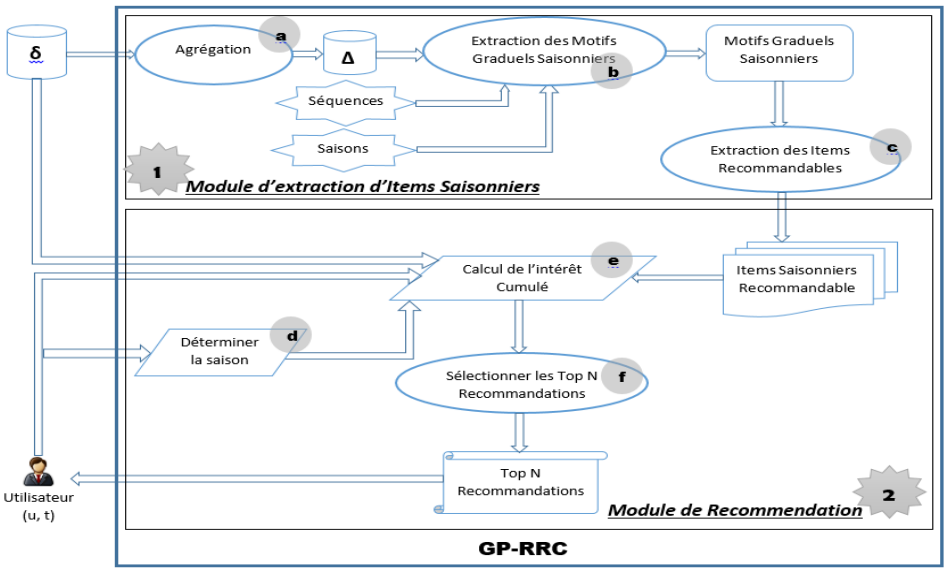

FIG. 1 – Architecture de l'approche GP-RRC.

La figure 1 décrit les différentes étapes de l'approche GP-RRC qui sont regroupées en deux grandes phases : l'extraction des items saisonniers et la recommandation. Etant donnée une base de données numériques δ, GP-RRC se déroule comme suit :

1. Construire une nouvelle base de données en agrégeant les consommations de tous les utilisateurs par unité de temps (jour, semaine, mois,,...) (fonctionnalité $\langle$a$\rangle$ dans la figure 1) ;
2. Extraire les items saisonniers à partir des motifs graduels saisonniers extraits de la nouvelle base de données obtenue à l'étape précédente (fonctionnalités $\langle$b$\rangle$ et $\langle$c$\rangle$) ;

Pour recommander à un utilisateur u au temps t, il faut procéder de la manière suivante :

1. Déterminer la saison s correspondante au temps t (fonctionnalité $\langle$d$\rangle$).
2. Sélectionner les items de la saison s, calculer l'intérêt cumulé de ces items et les ordonner (fonctionnalité $\langle$e$\rangle$).
3. Sélectionner les top-N items pour l'utilisateur u (fonctionnalité $\langle$f$\rangle$).

3.1 Aggrégation des données

Ici, nous montrons comment construire la base de données numérique $\Delta = \mathcal{T} \times \mathcal{I}$ à partir de la base de données initiale $\delta = \mathcal{U} \times \mathcal{I} \times \mathcal{T}$ Les tuples de la base δ sont de la forme (u, i, t) modélisant une matrice à trois dimensions où à l'intersection de ces trois entités, se trouve une grandeur numérique x_{uit} qui représente la quantité d'item i achetée par l'utilisateur u au temps t. En se basant sur δ, une nouvelle base Δ est créée avec en colonne les items et en ligne les temps. Formellement, pour un item i et un temps t, $\Delta_{t,i}$ est calculé de la manière suivante :

$$\Delta_{t,i} = \sum_{u \in \mathcal{U}} x_{uit} \quad (1)$$

3.2 Extraction des items saisonniers

Les motifs graduels saisonniers furent introduits pour la première fois dans Lonlac et al. (2020). Dans cet article, nous assouplissons la définition des motifs graduels saisonniers proposée en les considérant comme des motifs graduels qui apparaissent fréquemment dans la même saison (une séquence de périodes de temps) et pas nécessairement avec la même extension (Lonlac et al., 2020) la sous-séquence d'une saison. Pour ce faire, nous définissons une nouvelle approche qui consiste à extraire les motifs graduels fréquents en utilisant l'algorithme T-GPatterns (Lonlac et al., 2018) et à déterminer leur saisonnalité en post-traitement. Le choix de l'algorithme T-GPatterns est dû au fait qu'il est le plus approprié pour l'extraction de motifs graduels fréquents sous contrainte temporelle. Le principe de l'algorithme d'extraction de ces items saisonniers est le suivant :

— Définir l'ensemble des saisons $\mathcal{S}$ et séquences $\mathcal{A}$. Cette étape peut être faite par un expert ou en utilisant le calendrier civil ;
— Pour chaque saison $S_k \in \mathcal{S}$ $(1 \leq k \leq |\mathcal{S}|)$, capturer les données de chaque séquence définie par $\Delta_{S_k}^{A_p}$ où $A_p \in \mathcal{A}$ $(1 \leq p \leq |\mathcal{A}|)$;
— Pour chaque donnée capturée, extraire les motifs graduels fréquents définis par $M_{S_k}^{A_p}$, puis calculer le discriminant [1] de chacun de ces motifs graduels.
— Supprimer les motifs graduels qui ne sont pas fréquents dans la séquence (non saisonniers), c'est-à-dire les motifs graduels dont le discriminant est inférieur au disciminant seuil ($Mindisc$).
— Calculer le support cumulatif des motifs graduels restants en utilisant la formule 2.

$$SuppCum(m, M_{S_k}^*) = \frac{nbOcc(m, M_{S_k}^*) * \sum_{p=1}^{|\mathcal{A}|} Supp(m, M_{S_k}^{A_p})}{|\mathcal{A}|} \tag{2}$$

où $nbOcc(m, M_{S_k}^*)$ est le nombre de fois que le motif graduel m a été extrait et $M_{S_k}^*$ est l'union des motifs fréquents de chaque séquence pour la saison S_k $(1 \leq k \leq |\mathcal{S}|)$.
— Une fois les motifs graduels fréquents extraits pour toutes les saisons, discriminer ces derniers en fonction de leur support cumulatif. $\forall m \in M_{S_p}^*$ et $m \in M_{S_q}^*$ avec $p \neq q$, procéder comme suit :
  — Si $SuppCum(m, M_{S_p}^*) > SuppCum(m, M_{S_q}^*) \implies m \in M_{S_p}^*$;
  — Si $SuppCum(m, M_{S_p}^*) < SuppCum(m, M_{S_q}^*) \implies m \in M_{S_q}^*$;
  — Si $SuppCum(m, M_{S_p}^*) = SuppCum(m, M_{S_q}^*) \implies m \in M_{S_p s}^*$ and $m \in M_{S_q}^*$.
— Sélectionner les items de gradualité croissante. Cette action est faite parce que ces items de gradualité croissante sont ceux qui sont de plus en plus sollicités, alors que ceux de gradualité décroissante le sont de moins en moins. Pour cela, un poids associé à chaque item est calculé suivant la formule 3 où $\Omega^+ = |M_{S_k}^{i^+}| * \sum_{\forall m_p \in M_{S_k}^{i^+}} SuppCum(m_p, M_{S_k}^*)$ et $\Omega^- = |M_{S_k}^{i^-}| * \sum_{\forall m_q \in M_{S_k}^{i^-}} SuppCum(m_q, M_{S_k}^*)$

$$Cum(i, S_k) = \frac{\Omega^+ - \Omega^-}{|M_{S_k}|} \tag{3}$$

Où M_{S_k} est l'ensemble des motifs graduels fréquents de la saison S_k, $M_{S_k}^{i^+}$ (resp. $M_{S_k}^{i^-}$) l'ensemble des motifs graduels contenant l'item graduel $i^{\geq}$ (resp. $i^{\leq}$).

1. Nombre d'occurrences du motif graduel dans la saison divisé par le nombre de séquence

3.3 Recommandation d'items à un utilisateur

Pour recommander à un utilisateur donné u au temps t, après avoir déterminé la saison s correspondante au temps t, nous proposons de calculer l'intérêt cumulé ($CI(u,i)$ pour chaque item i obtenu des items saisonniers de la saison s. Soit $\mathcal{P}_u^t(i) = \{(i,t_1);(i,t_2);...,(i,t_k)\} \in \mathcal{P}_u^t$ avec $1 \leq k \leq N_u$, la séquence de consommation de l'item i par l'utilisateur u jusqu'au temps t. Cette grandeur (CI), calculée suivant l'équation 4 permet de savoir comment l'utilisateur u a interagi avec l'item i par le passé.

$$CI(u,i) = \sum_{q=2}^{k} \frac{1}{t_q - t_{q-1}} \times \sum_{(i,t)\in\mathcal{P}_u^t(i)} x_{uit} \tag{4}$$

4 Expérimentations

4.1 Description des données et environnement de travail

Nous présentons ici des expériences réalisées sur une base de données du monde réel d'un opérateur de tourisme et discutons des résultats. Cette base de données contient 593 utilisateurs et 409 items décrivant les voyages que ces utilisateurs ont effectués à travers le monde de 2012 à 2018 ; la quantité représente le nombre de voyages effectués par les utilisateurs. Nous avons expérimenté notre modèle en considérant deux scénarios : le scénario **One-To-One** qui étant donné une année y_1 pour évaluer les recommandations, considère uniquement les données de l'année précédente ($y_1 - 1$) comme données d'entraînement, et le scénario **One-To-Many** qui étant donné une année y_1 pour évaluer les recommandations, considère les données de toutes les années précédentes de y_1 comme données d'entraînement. Pour chaque scénario, nous avons considéré cinq valeurs de Minsupp ($Minsupp \in$0,2 ; 0,25 ; 0,3 ; 0,35 ; 0,4) et trois valeurs de Mindisc ($Mindisc \in \{50\%; 75\%; 100\%\}$). Lorsque Mindisc = 50% (resp. 75%), seule la moitié (resp. les trois quarts) des séquences extraites est prise en compte. Lorsque Mindisc = 100%, le nombre total de séquences extraites est utilisé.

Le modèle GP-RRC a été implémenté en R (version 3.6.0) pour la majorité des composants et Java Development Toolkits (JDK version 11.0.9) pour l'algorithme *AprioriTid* pour l'extraction des itemsets. Toutes les expérimentations ont été réalisées sur une machine de 15 Go de RAM, avec un processeur IntelR Core (TM) i5-8250U.

Nous utilisons la précision, le rappel et le gain cumulé actualisé normalisé (NDCG) (Pinel-Sauvagnat et Mothe, 2013) comme métriques pour la recommandation des Top-N.

4.2 Résultats

Nous avons divisé le jeu de données en saisons. D'après nos expérimentations, le nombre de saisons optimal dans la base de données "Tour Operator" est de quatre ; c'est-à-dire que chaque saison est constituée de 3 mois débutant du 1 Janvier au 31 Décembre. Nous avons également étudié l'impact du support sur la qualité de la recommandation et nous avons fait le constat que l'approche proposée donne de meilleurs résultats lorsque le support est bas. Nous allons considérer pour la suite le Minsupp = 0,2.

La table 1 présente les résultats obtenus lorsque le support est 0,2. Nous observons de cette table que les deux scénarios One-To-One et One-To-Many donnent de meilleurs résultats

quelque soit la valeur du Mindisc par rapport à SLRC. Nous déduisons de ces résultats que lorsque le support et le discriminant sont bien choisis, le modèle proposé donne de meilleurs résultats par rapport au modèle SLRC.

Algorithmes	Mindisc	Top N = 5			Top N = 10		
		Précision	Rappel	NDCG	Précision	Rappel	NDCG
One-To-Many	50%	0,11	0,43	0,52	0,21	0,42	0,56
One-To-Many	75%	0,11	0,43	0,52	0,21	0,42	0,56
One-To-Many	100%	0,11	0,43	0,52	0,21	0,42	0,56
One-To-One	50%	0,09	0,35	0,46	0,17	0,34	0,51
One-To-One	75%	0,11	0,31	0,45	0,20	0,30	0,51
One-To-One	100%	0,13	0,28	0,41	0,15	0,21	0,41
SLRC	/	0,06	0,05	0,04	0,05	0,11	0,07

TAB. 1 – Résultats comparatifs avec un support de 0,2.

5 Conclusion

Les interactions entre utilisateurs et items produisent souvent des grandeurs numériques (quantité d'items achetée, la note attribuée à un item par un utilisateur...) qui varient en fonction du temps et de l'utilisateur. La prise en compte de ces grandeurs numériques dans la recommandation de la consommation répétée est un grand défi. Dans cet article, nous proposons un modèle nommé GP-RRC basé sur les motifs graduels pour capturer efficacement les variations des interactions utilisateur-item (quantité) à travers le temps dans un contexte de la consommation répétée. Les expérimentations approfondies sur un jeu de données du monde réel confirment l'efficacité de la méthode proposée par rapport à l'état de l'art. Les travaux futurs incluent l'extension de GP-RRC à l'aide d'une approche collaborative qui consiste à prendre en compte les liens entre les utilisateurs en les regroupant par communauté. Il convient également de mener des expérimentations sur d'autres jeux de données plus volumineux.

Remerciements

Nous remercions les relecteurs anonymes pour leurs remarques constructives, et William Guyot (LIMOS) pour son aide durant la phase expérimentale.

Références

Anderson, A., R. Kumar, A. Tomkins, et S. Vassilvitskii (2014). The dynamics of repeat consumption. In *WWW*, pp. 419–430.

Benson, A. R., R. Kumar, et A. Tomkins (2016). Modeling user consumption sequences. In *WWW*, pp. 519–529.

Bhagat, R., S. Muralidharan, A. Lobzhanidze, et S. Vishwanath (2018). Buy it again : Modeling repeat purchase recommendations. In *ACM SIGKDD*, pp. 62–70.

Chen, J., C. Wang, J. Wang, et S. Y. Philip (2016). Recommendation for repeat consumption from user implicit feedback. *IEEE Transactions on Knowledge and Data Engineering 28*(11), 3083–3097.

Di-Jorio, L., A. Laurent, et M. Teisseire (2009). Mining frequent gradual itemsets from large databases. In *IDA*, Volume 5772, pp. 297–308.

Kapoor, K., K. Subbian, J. Srivastava, et P. Schrater (2015). Just in time recommendations : Modeling the dynamics of boredom in activity streams. In *WSDM*, pp. 233–242.

Lonjarret, C., R. Auburtin, C. Robardet, et M. Plantevit (2020). Sequential recommendation with metric models based on frequent sequences. *CoRR abs/2008.05587*.

Lonlac, J., A. Doniec, M. Lujak, et S. Lecoeuche (2020). Mining frequent seasonal gradual patterns. In *DaWaK*, Volume 12393, pp. 197–207.

Lonlac, J. et E. Mephu Nguifo (2020). A novel algorithm for searching frequent gradual patterns from an ordered data set. *Intell. Data Anal. 24*(5), 1029–1042.

Lonlac, J., Y. Miras, A. Beauger, V. Mazenod, J.-L. Peiry, et E. Mephu Nguifo (2018). An approach for extracting frequent (closed) gradual patterns under temporal constraint. In *FUZZ-IEEE*, pp. 1–8.

Parra, D. et S. Sahebi (2013). Recommender systems : Sources of knowledge and evaluation metrics. In *Advanced techniques in web intelligence-2*, pp. 149–175. Springer.

Pinel-Sauvagnat, K. et J. Mothe (2013). Mesures de la qualité des systèmes de recherche d'information. *Ingénierie des Systèmes d'Information. 18*(3), 11–38.

Schafer, J. B., J. Konstan, et J. Riedl (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pp. 158–166. ACM.

Teevan, J., E. Adar, R. Jones, et M. A. Potts (2007). Information re-retrieval : repeat queries in yahoo's logs. In *ACM SIGIR*, pp. 151–158.

Wang, C., M. Zhang, W. Ma, Y. Liu, et S. Ma (2019). Modeling item-specific temporal dynamics of repeat consumption for recommender systems. In *WWW*, pp. 1977–1987.

Summary

Recommendation systems were designed to solve the problem of data overload. The objective is therefore to select from a large number of items those of low quantity relevant to a given user. Taking into account the repetitive and periodic nature of interactions between users and items has improved the performance of existing systems. But these systems do not take into account the digital data associated with these interactions. In this paper, we propose a recommendation approach based on gradual patterns which makes it possible to model the covariations between items. The experimental results obtained with proposed approach are encoraging with respect to the used dataset.