SORBONNE UNIVERSITE
SCIENCE ET TECHNOLOGIES

PLDAC

Système de Recommandation Musicale

Yue WANG
Yuksel Selen ADALI

Février 2018

Table des Matières

Liste des Figures

1 REVUE DE LA LITTERATURE

1.1 DEFINITION DU PROJET

Dans le cadre d'un partenariat avec la société HorseCom, l'équipe MLIA dispose d'une base de données concernant l'entraînement des chevaux. Ils ont démontré l'intérêt de faire écouter de la musique aux chevaux (et aux cavaliers) pour améliorer la capacité de concentration des chevaux et leur régularité (rythmique). Tous les couples (cavalier,cheval) ne réagissent pas de la même façons aux différentes musiques: nous avons donc besoin d'apprendre un profil pour les différents couples -dans différents contextes, pour différents exercices- afin de proposer les morceaux les plus pertinents. Comme dans beaucoup d'application de recommandation, les propositions devront respecter une certaine diversité. Des modèles basés sur le filtrage collaboratif seront développés et combinés avec des stratégies diverses pour le démarage à froid (lorsqu'un nouvel utilisateur arrive dans le système). [*UPMC - PLDAC Sujets 2018* n.d.]

1.2 CONTEXTE GÉNÉRALE

Les systèmes de recommandations sont des outils logiciels et des techniques fournissant des suggestions d'éléments à utiliser par un utilisateur qui sont cruciaux pour mettre les profils d'utilisateurs en relation avec leurs correspondances de différentes manières.En conséquence, diverses techniques de génération de recommandations ont été proposées et, au cours de la dernière décennie, bon nombre d'entre elles ont également été déployées avec succès dans des environnements commerciaux.

Les recommandations personnalisées sont proposées sous forme de listes d'articles classifiés. En effectuant le ranking, les systèmes de recommandations tentent de prédire quels sont les produits ou services les plus appropriés, en fonction des préférences et des contraintes de l'utilisateur. Pour mener à bien une telle tâche de calcul, les systèmes de recommandation recueillent auprès des utilisateurs leurs préférences, qui sont soit explicitement exprimées. L'idée principal est d'améliorer ces profils d'utilisateur afin

d'augmenter le taux de pertinance des prédictions.[Ricci, Rokach, and Shapira, 2011]

TODO: raisons d'utilisation de systèmes recommandation, exemples des grands entreprises, types/ap

Ces systèmes de traitement def l'information qui recueillent activement divers types de données afin d'élaborer leurs recommandations. Dans cette étude, nous allons commencer à implementer nos méthodes de récommandation en utilisant la base de données de MovieLens et TODO: continuer avec la base de données de HorsCom. En partant de matrice rating basée sur les differents items et utilisateurs, nous allons essayer de completer les données manquants et augmenter la precision de nos recommandations.

1.3 LES DATASETS

Afin de développer notre système de recommandation nous avons utilisé la base de données MovieLens Small avec 100 000 classements et 1300 tags appliquées à 9 000 films par 700 utilisateurs. Les jeux de données MovieLens, publiés pour la première fois en 1998, décrivent les préférences exprimées par les gens pour les films. Ces préférences se présentent sous la forme de tuples: <utilisateur, item, cotation, horodatage>, chacun étant le résultat d'une personne exprimant une préférence (de 0 à 5 étoiles) pour un film à un moment donné afin de recevoir des recommandations personnalisées. [Harper and Konstan, 2016]

Nous avons créé une matrice rating $R_{u,i}$ correspondant aux utilisateurs et items(films) en séparant 90% pour l'entrainement et 10% pour le test. $R_{u,i}$ représente la note que utilisateur u donne à film i.

Les ***item***s sont les objets recommandés qu'ils peuvent être représentés en utilisant diverses approches d'information et de représentation et caractérisés par leur complexité et leur valeur ou leur utilité. Dans cet article, nous n'avons que des valeurs positives pour les items.

Les ***utilisateur***s d'un système de recommandations peuvent avoir des objectifs et des caractéristiques divers. Le choix des informations à modéliser dépend de la technique de recommandation. Par exemple, dans le filtrage collaboratif, les utilisateurs sont mod-

élisés comme une simple liste contenant les évaluations fournies par l'utilisateur pour certains éléments. Le modèle utilisateur jouera toujours un rôle central dans une approche de filtrage collaboratif, l'utilisateur est soit directement classé en fonction de ses ratings aux éléments, soit, à l'aide de ces ratings, le système dérive un vecteur de valeurs factorielles, où les utilisateurs diffèrent dans la pondération de chaque facteur dans leur modèle.

Les ***rating***s sont la forme de données transactionnelles la plus populaire qu'un système de recommandation recueille. Dans le recueil explicite des ratings, l'utilisateur est invité à donner son avis sur un élément sur une échelle de notation. Dans cette étude nous avons des évaluations numériques entre 0-5.

[TODO: Data-HorseCom]

2 METHODES ET TECHNIQUES

Kullanılması planlanan teknolojiler ve metotların kısa özeti - en az 1 sayfa, en fazla 2 sayfa.

Filtrage collaboratif

Dans cette approche, le système recommande à l'utilisateur actif les articles que d'autres utilisateurs avec des goûts similaires aimaient dans le passé. La similarité des goûts de deux utilisateurs est calculée sur la base de la similarité de l'historique de rating des utilisateurs. C'est la raison pour laquelle nous appellons filtrage collaboratif "corrélation entre les personnes". Le filtrage collaboratif est la technique la plus populaire et la plus répandue dans le système de recommandation. [Ricci, Rokach, and Shapira, 2011]

2.0.1 PROBLEME DES DONNEES MANQUANTES - REDUCTION DE LA DIMENTIONNALITE

Il est courant en système recommandation d'avoir non seulement un ensemble de données avec des caractéristiques qui définissent un espace à haute dimension, mais aussi des informations très rares dans cet espace (c'est-à-dire qu'il y a des valeurs pour un nombre limité de caractéristiques par objet).

Les techniques de réduction de la dimensionnalité aident à surmonter ce problème en transformant l'espace original à haute dimension en une dimension plus basse.

La rareté et la malédiction de la dimensionnalité sont des problèmes récurrents en système recommandation. Même dans le cadre le plus simple, nous avons probablement une matrice sparse avec des milliers de lignes et de colonnes (utilisateurs et articles), dont la plupart sont des zéros. Par conséquent, la réduction de la dimensionnalité intervient naturellement.

2.0.1.1 Décomposition en Valeur Singulière

L'application de la décomposition en valeur singulière, dans le domaine du filtrage collaboratif nécessite de factoriser la matrice rating utilisateur-item. Cela soulève souvent des difficultés en raison de la proportion élevée de valeurs manquantes causées par la rareté dans la matrice d'évaluation des articles de l'utilisateur. Décomposition en valeur singulière conventionnel n'est pas défini lorsque les connaissances sur la matrice sont incomplètes.

La décomposition en valeurs singulières est indéfinie dans le cas de matrices incomplètes. Nous allons donc utiliser d'autres algorithmes d'apprentissage.[]

2.0.1.2 Méthode de Factorisation en Matrices Non-Négatives

TODO: nmf

La plupart des classificateurs et des techniques de regroupement dépendent fortement de la définition d'une similarité ou d'une mesure de distance appropriée. Pour trouver une factorisation approximative $V \approx W \cdot H$ il faut d'abord définir des fonctions de coût qui quantifient la qualité de l'approximation. Une telle fonction de coût peut être construite en utilisant une certaine mesure de la distance entre deux matrices non négatives A et B.

L'exemple le plus simple et le plus commun d'une mesure de distance est la distance euclidienne:

$$\|A-B\|^2 = \sum_{ij}(A_{i,j}-B_{i,j})^2$$

Une autre mesure utile réduit à la divergence de Kullback-Leibler:

$$D(A||B) = \sum_{ij}(A_{i,j}\cdot \log\frac{A_{i,j}}{B_{i,j}} - A_{i,j} + B_{i,j})$$

Comme la distance euclidienne, cette mesure est également disparaît si et seulement si A = B. Mais on ne peut pas l'appeler "distance", parce qu'elle n'est pas symétrique pour

A et B, c'est pourquoi nous l'appelons la "divergence" de A par rapport à B. [Lee and Seung, 2001]

Afin de surmonter les risques du sur apprentissage dans le cas ou la matrice d'entrée est très sparse, nous allons faire la régularisation.

Les systèmes antérieurs s'appuyaient sur l'imputation pour remplir les valeurs manquantes et densifier la matrice de cotation. Cependant, l'imputation peut être très coûteuse car elle augmente considérablement la quantité de données. En outre, une imputation inexacte peut fausser considérablement les données. Par conséquent, des travaux plus récents ont suggéré de modéliser directement les notations observées seulement, tout en évitant le dépassement par un modèle régularisé. Pour apprendre les vecteurs factoriels, le système minimise l'erreur moindre carrées sur l'ensemble des notations connues:

$$\arg\min \sum_{iu} (u_i - r_{ui})^2 + (\lambda_u \cdot \|u\|^2 + \lambda_i \cdot \|i\|^2)$$

[Bell, Koren, and Volinsky, 2008]

2.0.2 DESCENTE DE GRADIENT STOCHASTIQUE

3 DEROULEMENT DU PROJET

BIBLIOGRAPHIE

Bell, R. M., Koren, Y., and Volinsky, C. (2008). The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research*. 1.

Harper, F. M. and Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*. 5.4, p. 19.

Lee, D. D. and Seung, H. S. (2001). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems*, pp. 556–562.

Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In: *Recommender systems handbook*. Springer, pp. 1–35.

UPMC - PLDAC Sujets 2018. `http://dac.lip6.fr/master/wp-content/uploads/2017/12/guigue_gallinari.pdf`. Sorbonne Université, Master Informatique DAC: Sujets de PLDAC http://dac.lip6.fr/master/enseignement/ues/pldac/sujets-pldac-2018/.