

Pour une meilleure exploitation de la classification croisée dans les systèmes de filtrage collaboratif

Aghiles Salah, Nicoleta Rogovschi, François Role, Mohamed Nadif

LIPADE, Université Paris Descartes, France
45, rue des Saints Pères,
75006, Paris
Prénom.Nom@parisdescartes.fr

Résumé. Pour la prédiction automatique des items préférés par des utilisateurs sur le Web, différents systèmes de filtrage collaboratif ont été proposés. La plupart d'entre eux sont basés sur la factorisation matricielle et les approches de type k plus proches voisins. Malheureusement ces deux approches requièrent un temps de calcul important. Une partie de ces problèmes a pu être surmontée par la classification croisée ou *co-clustering* qui s'avère pertinente du fait qu'elle permet par nature une gestion simultanée des ensembles correspondant aux utilisateurs et aux items. Cependant, des travaux doivent encore être menés pour une meilleure prise en compte des données manquantes. Dans ce travail, nous proposons donc une gestion efficace des données non observées permettant une meilleure exploitation du potentiel de la classification croisée dans le domaine des systèmes de recommandation. Nous montrons de plus qu'elle permet d'obtenir des représentations à base de graphes bipartis facilitant l'interprétation interactive des affinités entre des groupes d'utilisateurs et des groupes d'items.

1 Introduction

L'objectif des systèmes de recommandation est de prédire les choix et les préférences individuelles en fonction des comportements et des préférences observées. Le filtrage collaboratif est la technique la plus utilisée par les systèmes de recommandation. Il consiste à comparer les données d'un utilisateur avec des données similaires d'autres utilisateurs, basée sur les habitudes d'achat et de navigation (Goldberg et al., 1992). Il permet aux commerçants de fournir des recommandations aux clients pour de futurs achats. Dans la suite, les données sont représentées par une matrice \mathbf{U} de taille $(n \times p)$ où chaque ligne représente un utilisateur, les colonnes représentent des items, et chaque cellule (u_{ij}) de \mathbf{U} est la note attribuée par un utilisateur i pour un item j . Les notes (u_{ij}) peuvent être binaires, ou réelles et dans ce cas \mathbf{U} est appelée matrice réelle de notations. La matrice \mathbf{U} peut être obtenue de manière explicite (en gardant les évaluations fournies par les utilisateurs pour des articles donnés) ou de manière implicite (en considérant qu'un utilisateur préfère implicitement acheter ou pas les éléments présentés sur des pages Web visitées).

Dans le filtrage collaboratif (désormais désigné par FC), plusieurs approches sont utilisées. Les techniques de FC actuelles telles que celles basées sur la corrélation entre utilisateurs