Apport des données thématiques dans les systèmes de recommandation : hybridation et démarrage à froid

Frank Meyer*, Eric Gaussier**
Fabrice Clerot*, Julien Schluth***

* Orange Labs, 2, Av. Pierre Marzin 22307 Lannion cedex
franck.meyer@orange-ftgroup.com, fabrice.clerot@orange-ftgroup.com
** Laboratoire d'Informatique de Grenoble BP 53 - 38041 Grenoble cedex 9
Eric.Gaussier@imag.fr
*** julien.schluth@gmail.com

Résumé. Des travaux récents (Pilaszy et al., 2009) suggèrent que les métadonnées sont quasiment inutiles pour les systèmes de recommandation, y compris en situation de cold-start : les données de logs de notation sont beaucoup plus informatives. Nous étudions, sur une base de référence de logs d'usages pour la recommandation automatique de DVD (Netflix), les performances de systèmes de recommandation basés sur des sources de données collaboratives, thématiques et hybrides en situation de démarrage à froid (cold-start). Nous exhibons des cas expérimentaux où les métadonnées apportent plus que les données de logs d'usage (collaboratives) pour la performance prédictive. Pour gérer le cold-start d'un système de recommandation, nous montrons que des approches "en cascade", thématiques puis hybrides, puis collaboratives, seraient plus appropriées.

1 Introduction

Le but des systèmes de recommandation automatique est d'aider des utilisateurs à trouver des produits (appelés items) qui les intéressent dans de très grands catalogues. On distingue en général deux types de technique, celle dite de filtrage collaboratif, et celle dite de filtrage thématique (Adomavicius & Tuzhilin, 2005). Lorsqu'un système de recommandation utilise à la fois des données d'usages des "utilisateurs" et des métadonnées des "items", on parle de système de recommandation hybride (Burke, 2007). Les systèmes de recommandation ont un point faible connu, celui dit du cold start : en l'absence de données d'usages, leurs performances sont détériorées (Adomavicius & Tuzhilin, 2005). Dès lors, utiliser des approches dites thématiques, c'est-à-dire basées sur des métadonnées descriptives des items, à la place des méthodes collaboratives, ou en renfort, est une idée récurrente. Les méthodes thématiques et hybrides ont été à notre connaissance peu étudiées sur les données de référence largement accessibles que constitue la base Netflix (Netflix Prize 2007). (Pilazy et al., 2009) abordent en partie ce sujet en utilisant des méthodes de factorisation de matrices appliquées à des données de Netflix aussi bien collaboratives (logs du Netflix Prize), que thématiques (site www.netflix.com). Mais leur étude aboutit à la conclusion que les métadonnées présentent peu d'intérêt. Nous étudions pour notre part l'apport des méthodes thématiques et hybrides selon une technique de type k-plus-proches voisins, qui est plus utilisée en contexte opérationnel (Koren, 2010), et selon un protocole dédié à l'analyse du cold start..