Analyse exploratoire d'opinions cinématographiques : co-clustering de corpus textuels communautaires

Damien Poirier*, Cécile Bothorel*
Marc Boullé*

*TECH / EASY
France Telecom RD
2 avenue Pierre Marzin
22300 Lannion
prénom.nom@orange-ftgroup.com,
http://www.francetelecom.com/fr/groupe/rd/

Résumé. Les sites communautaires sont un endroit privilégié pour s'exprimer et publier des opinions. Le site *www.flixster.com* est un exemple de site participatif sur lequel se rassemblent plus de 20 millions de cinéphiles qui partagent des commentaires sur les films qu'ils ont ou non aimés. Explorer les contenus autoproduits est un challenge pour qui veut comprendre les attentes des internautes. Par une méthode d'apprentissage non supervisée, nous montrerons qu'il est possible de mieux comprendre le vocabulaire utilisé pour décrire des opinions. En particulier, grâce à une méthode de co-clustering, nous montrerons qu'un rapprochement peut être fait entre des films particuliers sur la base de l'usage d'un vocabulaire particulier. L'analyse des résultats peut conduire à retrouver une certaine typologie de films ou encore des rapprochements entre films. Cette étude peut être complémentaire avec des analyses linguistiques des corpus, ou encore être exploitée dans un contexte applicatif de recommandation de contenus multimédias.

1 Introduction

Les avancées technologiques en matière de haut débit favorisent l'apparition de nouveaux services de vente ou location en ligne de fichiers vidéos et musicaux. De tels services se veulent pro-actifs et proposent, en plus des actes promotionnels classiques, des choix personnalisés de films (ou de musique). Des méthodes de recommandation sont déjà utilisées sur certains sites Internet de vente par correspondance (*Amazon*, *Fnac*, *Virgin*, etc.) ou encore sur les plates-formes musicales (*Lastfm*, *Radioblog*, *Pandora*, etc.). Candillier et al. (2007) fait un panorama des techniques de recommandation : qu'elles soient basées sur des notations d'internautes ou des descriptions de contenus (techniques *user-* and *item-based* utilisant le filtrage collaboratif) ou des rapprochements thématiques de profils d'internautes et de descriptions de contenus (filtrage de contenus), voire des techniques hybrides combinant les différentes approches, la problématique reste de gérer les *matrices creuses*. En effet, devant la variété d'un catalogue et le grand nombre d'utilisateurs, le faible nombre de notes qu'un utilisateur donne rend la