AbsTop-K α: un algorithme d'extraction de paires abstraites hautement corrélées pour mieux recommander dans la "longue traine"

Minh Thu TRAN NGUYEN[*,**,***] , François SEMPE[***]
Jean Daniel ZUCKER[*,**,****]

[*]LIMBIO,UFR SMBH Léonard de Vinci, Université Paris 13, 93017 Bobigny France
[**]IRD UMI 209,Centre IRD France Nord, Bondy, F-93143 France
[***]MSI, Institut de la francophonie pour l'Informatique (IFI), Ha Noi, Vietnam
[****]UPMC Université Paris 6, UMI 209, Paris, F-75006, France

Résumé. De nombreux systèmes de recommandation se focalisent sur les articles (que nous appellerons "items") les plus "populaires" et ignorent souvent la "longue traîne" des produits qui le sont moins. Nous proposons l'algorithme AbsTop-kα qui améliore les recommandations en se basant sur la combinaison (pondérée par α) de paires hautement corrélées entre des abstractions d'items et entre des paires d'items concrets classiquement recherchées.

1 Introduction

La problématique des systèmes de recommandation d'e-commerce (Dias et al., 2008), peut se formuler de manière simplifiée comme un problème de prédiction où, connaissant p items $X_1 \ldots X_p$ déjà choisis par un utilisateur U il s'agit de proposer k items cibles ayant le plus de chance d'être choisis par l'utilisateur. Dans la pratique p est faible et k vaut souvent trois. De fait, la distribution des items dans les transactions possède de "longues traînes", suivant ainsi une loi dite de Pareto des "80/20". En d'autres termes, pour tout item Xj potentiellement recommendable, on cherche la probabilité $P(X_j|X_i, \ldots, X_p, U)$(Dias et al., 2008). Cet article propose une approche basée sur la recherche de paires hautement corrélées entre des items abstraits pour améliorer les performances vis-à-vis du problème de la "longue traîne" et en partie de celui du "cold start".

2 L'algorithme AbsTop-Kα et expérimentations

L'intuition au coeur de l'approche est dans la définition d'items abstraits qui regroupent des items concrets et la recherche de règles d'associations abstraites entre des paires d'items abstraits. Le recours à une abstraction sur les items a été proposé par Han et Fu mais dans le cadre de la recherche de règles d'associations de granularités différentes (Han et al., 1999). L'algorithme AbsTop-Kα donne la liste des K items (dans la pratique on prends souvent K=3 (Sordo-Garcia et al., 2007)) provenant pour K$*(1-\alpha)$ items de la recommandation concrète

(par les paires de XiXj) et K$*\alpha$ paires issues de la recommandation abstraite (par les paires de AiAj). La proportion de recommandations abstraites est donnée par α . Ainsi pour α =33% on a $K = 3*1/3 = 1$ recommandation issue de règles abstraites (33% de 3) entre des abstractions d'items (et donc 2 issues de règles concrètes). Pour expérimenter les recommandations, comme la plupart des auteurs, nous avons eu recours à une base de données réelle qu'il n'est pas possible de rendre public pour des raisons de confidentialité. Nous avons utilisé les données de 2007 (9332 items ; 451 items abstraits ; 30009 transactions) pour calibrer AbsTop-Kα et celles de janvier et février 2008 (3943 items ; 365 items abstraits ; 5097 transactions) pour l'évaluer. Nous avons utilisé les procédures Given-N et AllButOne proposées par Hsu. et al. (Hsu et al., 2004) qui ont le mérite d'être rigoureuse et de quantifier l'amélioration de la qualité des recommandations de sites marchands.

α	Gain	# Bonnes Recos	# Nouveau Produits	# Longues traînes
0	0	541	0	262
1/3	21%	656	7,62%	312
2/3	24%	669	11,55%	314
1	−12%	477	18,45%	179

TAB. 1 – *Résultats de l'évaluation d'AbsTop-Kα pour K=3 et la valeur de α variant de 0 à 1.*

Les recommandations qui sont issues de l'algorithme AbsTop-Kα sont meilleures pour les α supérieur à 0 et inférieur à 1 (c.a.d. quand il y a des combinaison entre des recommandation abstraites et des recommandation concrètes). Le gain qualitatif en terme de nombre absolu de recommandations dans la longue traîne et les nouveaux items est significatif.

Les auteurs tiennent à remercier Edith Nuss, Hui Xiong et Wenjun Zhou auteurs de l'algorithme TOP-COP ainsi que l'ANR Compalimage qui a partiellement financé ces recherches.

Références

Dias, M. B., D. Locher, M. Li, W. El-Deredy, et P. J. Lisboa (2008). The value of personalised recommender systems to e-business : a case study. pp. 291–294.

Han, J., Y. Fu, I. C. Society, et I. C. Society (1999). Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering 11*, 798–805.

Hsu, C.-N., H.-H. Chung, et H.-S. Huang (2004). Mining skewed and sparse transaction data for personalized shopping recommendation. *Mach. Learn. 57*(1-2), 35–59.

Sordo-Garcia, C. M. et al. (2007). Evaluating retail recommender systems via retrospective data : Lessons learnt from a live-intervention study. pp. 197–206.

Summary

Many recommandation Systems only rely on the most popular items and ignore the long tail of items that are either less popular or new ones. We propose the algorithm AbsTop-Kα, where α balances between highly correlated pairs of abstract and concrete items.