

Algorithmes pour les systèmes de recommandation : un comparatif

Diane Gallois-Wong, Raphaël Rieu-Helft

27 janvier 2015

Introduction

Collaborative filtering

Objectif :

- Requêtes rapides
- Précision
- Ajout rapide d'un nouvel utilisateur
- Démarrage à froid

Algorithmes témoins

- Witness
- UnbiasedWitness
- PerUserAverage
- BiasFromMean

Slope One

$$dev_{j,k} = \sum_{i \in S^{j,k}} \frac{r_{i,j} - r_{i,k}}{\text{card} S^{j,k}}$$

$$\tilde{r}_{i,j} = \bar{u}_i + \frac{1}{\text{card} R_{i,j}} \sum_{k \in R_{i,j}} dev_{j,k}$$

- Calculs simples
- Réponse rapide à une requête $\tilde{r}_{i,j}$? (précalcul de dev)
- Ajout d'un nouvel utilisateur : actualisation facile

SVD

Variations sur la question bonus du DM2

- ShiftSVD
- UnbiasedSVD

Similarité cosinus

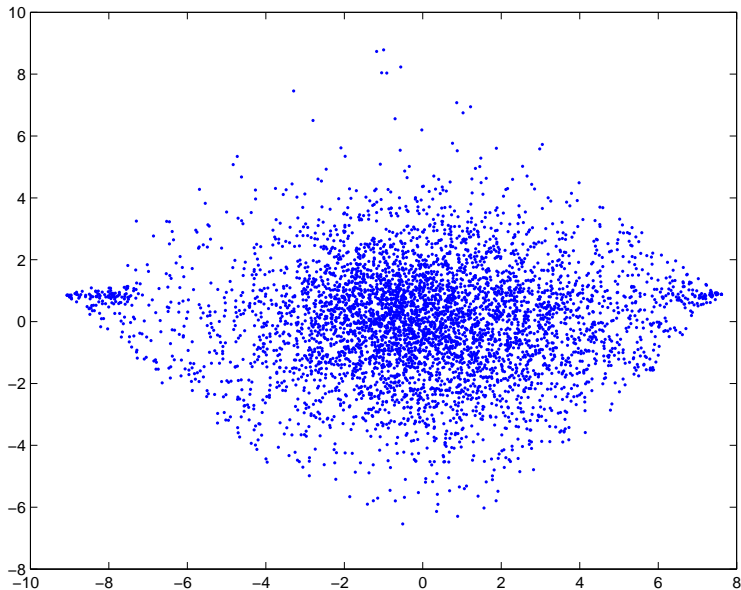
Algorithmes du cours : gestion des biais puis moyenne pondérée par un score de similarité

- UserCosSim
- ItemCosSim

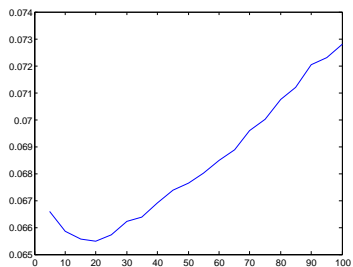
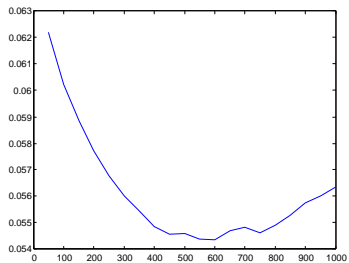
Analyse en composantes principales : Eigentaste

- Nécessite un petit ensemble d'objets notés par tout le monde
- Projection des utilisateurs sur un sous-espace bien choisi
- Clustering, puis méthode des histogrammes

Projection des vecteurs des utilisateurs sur un plan



Choix du nombre de clusters



Erreur MSE (norme 2)

Algorithme	bonus, p=0.05	Jester, p=0.05	bonus, p=0.2	Jester, p=0.2	bonus, p=0.5
Witness	0.0176	0.0673	0.0149	0.0145	0.0094
UnbiasedWitness	0.0126	0.0414	0.0085	0.0081	0.0051
PerUserAverage	0.0180	0.0471	0.0149	0.0089	0.0093
BiasFromMean	NaN	0.0417	0.0086	0.0081	0.0051
PlainSVD	0.1618	0.1839	0.0943	0.1737	0.0338
ShiftSVD	0.0191	0.0850	0.0106	0.0748	0.0050
UnbiasedSVD	0.0128	0.0683	0.0086	0.0511	0.0048
UserCosSim	0.0125	0.0505	0.0093	0.0426	0.0079
ItemCosSim	NaN	0.0456	0.0092	0.0307	0.0069
SlopeOne	NaN	0.0425	0.0086	0.0081	0.0052
WeightedSlopeOne	NaN	0.0499	0.0085	0.0083	0.0052
BiPolarSlopeOne	NaN	0.0479	0.0093	0.0084	0.0058

Jeux de données :

- Matrice de la question bonus du DM
- Jester : ensemble de ratings d'une centaine de blagues par plus de 50000 utilisateurs, avec un sous-ensemble noté par tout le monde

Erreur MAE (un rating masqué à la fois, norme 1)

Algorithme	p=0.1	p=0.2
Witness	1.0992e-05	2.7715e-06
UnbiasedWitness	6.6646e-06	1.8463e-06
PerUserAverage	9.1804e-06	2.3797e-06
BiasFromMean	8.7602e-06	1.4253e-06
ShiftSVD	7.1533e-06	1.5175e-06
UnbiasedSVD	5.4096e-06	1.5922e-06
UserCosSim	4.7402e-06	1.4364e-06
ItemCosSim	6.6190e-06	1.6920e-06
SlopeOne	7.1207e-06	1.5497e-06
WeightedSlopeOne	5.9259e-06	1.8653e-06
BiPolarSlopeOne	6.6666e-06	2.0102e-06

Bilan

- Selon le jeu de données et l'objectif (norme d'erreur), le choix de l'algorithme est très variable
- Compromis à faire entre précision, temps de calcul, mémoire, vitesse des requêtes et des mises à jour

Références

K. Goldberg, T. Roeder, D. Gupta, C. Perkins :

Eigentaste : A constant time collaborative filtering algorithm

D. Lemire, A. Maclachlan :

Slope One Predictors for Online Rating-Based Collaborative Filtering