Collaborative Filtering

Virgil Pavlu November 3, 2008

1 Collaborative filtering

- setup of the problem
  - why is different
  - examples : Amazon, Netflix, iTunes genius

		Airplane	**Matrix**	**Room with a View**	**...**	**Hidalgo**
		comedy	**action**	**romance**	**...**	**action**
Joe	*27,M,70k*	9	7	2		7
Carol	*53,F,20k*	8		9		
...						
Kumar	*25,M,22k*	9	3			6
U_a	*48,M,81k*	4	7	?	?	?

2 Pearson's Formula

COLLABORATIVE FILTERING; PEARSON FORMULA
compute for each user u mean and variance. Let N_u = number of movies rated by user u; R_{um} is the rating of user u for movie m

$$\mu_u = \frac{\sum_m R_{um}}{N_u}$$

$$\sigma_u = \frac{\sum_m R_{um}^2}{N_u} - \mu_u^2$$

normalize each ratings by substracting the user mean and dividing by user variance

$$\bar{r}_{um} = \frac{R_{um} - \mu_u}{\sigma_u}$$

compute user similarity between any two users u and v

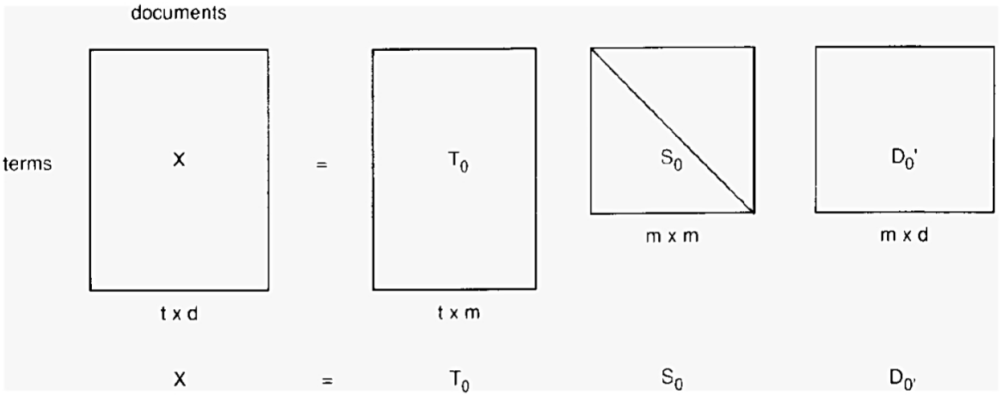
$$\rho_{uv} = \frac{1}{\text{movies in common } m} \sum_m \bar{r}_{um} \cdot \bar{r}_{vm}$$

predict the rating for a new movie by accounting for all other users' v rating on the movie

$$predict(u, m) = \mu_u + \frac{\sum_v \rho_{uv} \cdot \bar{r}_{vm}}{\sum_v |\rho_{uv}|} \cdot \sigma_u$$

3 kNN for Collaborative filtering

4 Factorization, LSI



- T_0, D_0 orthogonal with unit length columns
  - $T_0 * T_0^T = 1$
- S_0 = diagonal matrix of eigenvalues
- m = rank of X

41

LSI dimensions are
- based on term co-occurrence patterns between documents (profiles)
- ordered by their prominence in collection

LSI space built from profiles
- highlights common patterns among profiles
- "noisy" dimensions can be pruned
- project new documents into a collaborative space for routing

5 Netflix and recent advances

6 more examples

- technical paper recommendations
    - librarian tasks