Institute for Informatics, Ludwig-Maximilians-Universität (LMU) München, Germany

Institute of Computer Science, FORTH Hellas, Heraklion, Greece.

Exploring Subspace Clustering for Recommendations

User Based Collaborative Filtering

Recommendations for a user *u* are based on the ratings of his/her similar users (Friends F_u).

	1 (Titanic)	2 (Braveheart)	3 (Matrix)	4 (Inception)	5 (Hobbit)	6 (300)
Susan (Target user)	5	?	5	5	4	?
Bill	3	3	?	1	?	1
Jenny	5	4	1	1	?	4
Tim	?	?	4	4	3	3
Thomas	?	1	1	4	?	4

Similar users w.r.t. all movies

$$\hat{r}(u,i) = \frac{\sum_{u' \in F_u} sim(u,u') * r(u',i)}{\sum_{u' \in F_u} sim(u,u')}$$

Existing Approaches on Computing The Set Of Friends F_u

Nearest Neighbors

$$F_u = \{u' \in U : sim(u,u') \geq \delta\}$$

δ is the user similarity threshold

- Friends are defined in the full dimensional feature space
- Linear scan of the DB to compute F_u

Full Dimensional Clustering

To speed up friends computation, users are grouped into clusters of similar users $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$, $\theta_i \cap \theta_j = \emptyset$.

$$F^{clu}_u = \{u' \in \theta_i : u \in \theta_i\}$$

- Faster than nearest neighbors approach
- For each $u,u' \in \theta_i$, $sim(u,u') \geq \delta$ (*correctness*)
- $F^{clu}_u \subseteq F_u$ (*incompleteness*)
  - For small clusters, F^{clu}_u too narrow

Diversification Of The Set Of Friends

For both cases, user similarity is evaluated *in the full* (*high*) *dimensional* feature space

- Its difficult to find similar users when so many dimensions are considered

Its more probable for users to exhibit similarity *in some subspace* of the feature space

- e.g., similar taste in comedies but not in dramas

Subspace Clustering Based Recommendations

- Clusters are defined in subspaces of the original feature space: $\theta = (U_\theta, I_\theta)$
- Subspace clustering: $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$
  - θ_i, θ_j might overlap w.r.t. both users and items
- Fault tolerant subspace clustering for dealing with missing ratings
  - Missing values are tolerated but bounded within a cluster per user (ε_u), per item (ε_i) and in total (ε_g).

The friends of a user u are the members of all subspace clusters where u belongs to:

$$F^{subclu}_u = \{u' \in \theta_i : u \in \theta_i\}$$

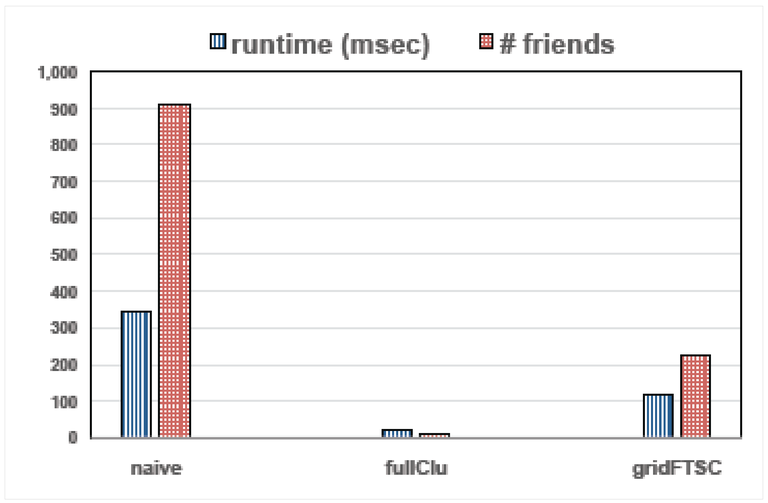
	1 (Titanic)	2 (Braveheart)	3 (Matrix)	4 (Inception)	5 (Hobbit)	6 (300)
Susan (Target user)	5	?	5	5	4	?
Bill	3	3	?	1	?	1
Jenny	5	4	1	1	?	4
Tim			4	4	3	3
Thomas	?	1	1	4	?	4

Similar users w.r.t. movies 1,2 — Similar users w.r.t. movies 3,4 — Similar users w.r.t. movies 5,6

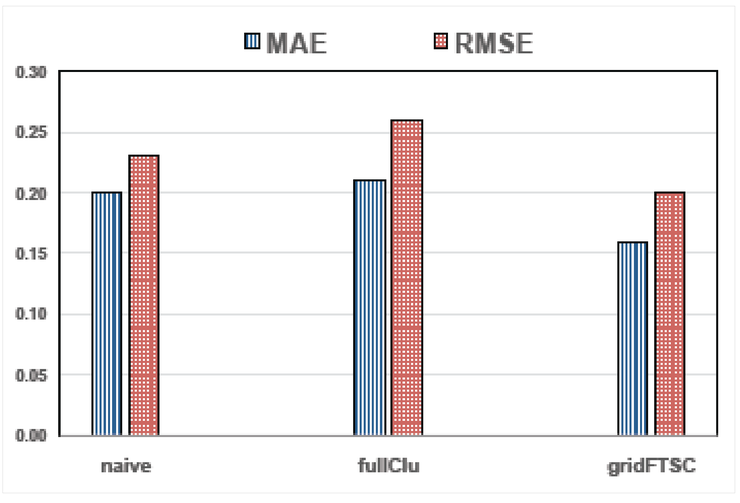
Refinement

Weighted Full Dimensional Ranking

- Rank the users in F^{subclu}_u based on their full dimensional distance to u
  - Weight by #shared dimensions
- Select the most prominent ones



Tradeoff in exec time & # friends



Better quality of recommendations

Katharina Rausch[1], Eirini Ntoutsi[1], Kostas Stefanidis[2], Hans-Peter Kriegel[1]

[1]Institute for Informatics, Ludwig-Maximilians-Universität (LMU) Munich, Germany

[2]Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science, Heraklion, Greece.