# Using Graph Partitioning Techniques for Neighbour Selection in User-Based Collaborative Filtering

Alejandro Bellogín[1], Javier Parapar[2]

*alejandro.bellogin@uam.es, javierparapar@udc.es*

[1]Information Retrieval Group, Department of Computer Science, Universidad Autónoma de Madrid

[2]Information Retrieval Lab, Department of Computer Science, University of A Coruña

UNIVERSIDAD AUTONOMA DE MADRID

UNIVERSIDADE DA CORUÑA

## Introduction

- Memory-based CF algorithms are based on the principle that a *particular user's rating records are not equally useful* to all other users as input to provide them with item suggestions [2].
- **Neighbourhood identification** is based on selecting those users who are most similar to the active user according to a certain similarity metric.
- **Claim:** cluster-based CF techniques may be improved by using spectral clustering methods, instead of old-fashioned clustering methods such as k-Means or hierarchical clustering.
- **Experiment:** a **spectral clustering** technique (*NCut*) has been introduced into a cluster-based CF algorithm which outperforms other standard techniques in terms of **ranking precision**.

## Normalised Cut for Collaborative Filtering

The outline of the spectral clustering algorithm using Normalised Cut would be as follows [3]:

1. Create a graph $G = (V, E, W)$ from the data collection to cluster the users, represented as the vertices $V$ in the graph $G$
2. Solve a trace minimisation problem involving the weights $W$ and the Laplacian matrix of the graph
3. The eigenvectors of the optimal matrix allow a projection of each data instance in $\mathbb{R}^k$ based on its similarity to the other instances
4. Cluster the resulting projected data points
5. Backtrace the grouping found in the previous step to the original data points, obtaining the final outcome of the clustering algorithm

In Collaborative Filtering, users play the role of the **nodes of the graph** to cut.

$$\tilde{r}(u,i) = \frac{\sum_{v \in NC(u)} \mathrm{sim}(u,v) r(v,i)}{\sum_{v \in NC(u)} |\mathrm{sim}(u,v)|}$$

- $NC(u)$ *outputs the elements who belong to the same cluster* as the target user $u$
- $\mathrm{sim}(u,v)$: similarity between users $u$ and $v$
- $r(v,i)$: rating given by user $v$ to item $i$

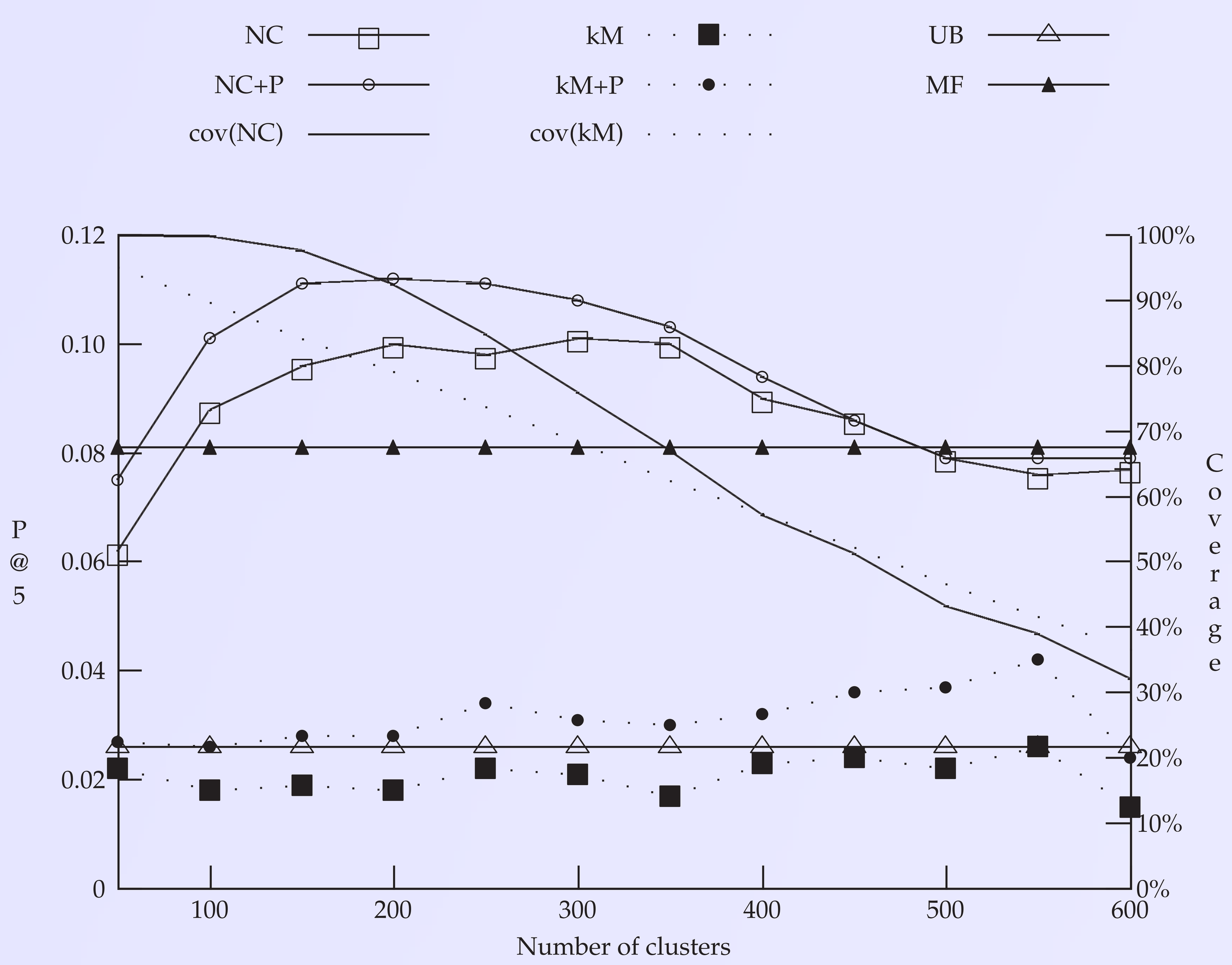
Notation: $\boxed{\text{NC+P when sim}(u,v)\text{ is Pearson}}$ and $\boxed{\text{NC when sim}(u,v) = 1}$

## Experiments and Results

**Evaluation methodology:** *TestItems* [1] (for each user a ranking is generated by predicting a score for every item in the test set).
**Baselines:** *UB* (user-based CF with Pearson's correlation as similarity measure), *MF* (matrix factorization algorithm with a latent space of dimension 50), *kM* and *kM+P* (user-based CF using k-Means clustering alone or with Pearson's similarity).
**Dataset:** *Movielens 100K*.

| Method | P@5 | Coverage |
|---|---|---|
| MF | $0.081^{u}$ | 100% |
| UB | 0.026 | 100% |
| NC+P 50 (100) | $0.075^{u}$ | 100.00% |
| NC+P 100 (150) | $0.101^{mu}$ | 97.82% |
| NC+P 150 (200) | $0.111^{mu}$ | 93.68% |
| NC+P 200 (250) | $\mathbf{0.112}^{mu}$ | 79.74% |
| NC+P 250 (300) | $0.111^{mu}$ | 69.06% |
| NC+P 300 (350) | $0.108^{mu}$ | 59.26% |
| NC+P 350 (400) | $0.103^{mu}$ | 54.25% |
| NC+P 400 (450) | $0.094^{mu}$ | 43.36% |
| NC+P 450 (500) | $0.086^{u}$ | 40.52% |
| NC+P 500 (550) | $0.079^{u}$ | 35.95% |
| NC+P 550 (600) | $0.079^{u}$ | 32.03% |
| NC+P 600 (650) | $0.079^{u}$ | 25.93% |

Performance results. In brackets, the number of eigenvectors used in the Normalised Cut for each cluster of size $k$.




Sensitivity of the performance and coverage to the number of clusters.

## Conclusions

- The use of $NCut$ as a clustering method (based on graph partitioning) for exploiting the neighbourhoods outperform other state-of-the-art approaches.
- The improvement in performance is consistent even when no similarity is used (method *NC*).
- The coverage of our method (measured as the number of users for which the system can recommend items) is higher than other clustering methods in similar conditions.

In the future, we aim to investigate also item clusters along with item-based approaches and alternative clustering methods.

## References

[1] Bellogín, A., Castells, P., and Cantador, I. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *RecSys* (2011), pp. 333–336.

[2] Desrosiers, C., and Karypis, G. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*. 2011, pp. 107–144.

[3] Shi, J., and Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 8 (2000), 888–905.

**RecSys 2012**, 6th ACM Conference on Recommender Systems. September 9-13, 2012. Dublin, Ireland.