# Cluster-based Collaborative Filtering Recommendation

**Roger H.L. Chiang**

**University of Cincinnati**

**October 24, 2008**

1

## Collaborator

Professor Chih-Ping Wei

College of Technology Management

National Tsing Hua University

Hsinchu, TAIWAN

2

## Outline

- Research Issue
- Motivation and Objectives
- Literature Review
- Cluster-based Collaborative Filtering Recommendation
- Empirical Evaluation
- Conclusions and Future Research

3

## Research Issue

- Internet establishes a digital marketplace (marketspace).
- Vast amount of product information is available online.
  - More information *should* support consumers make better purchase decisions.
  - However, **information overload/cognitive overload** becomes a critical challenge.
- Recommendation agents have emerged as e-service to:
  - address information/cognitive overload challenge for consumers, and
  - support customization and personalization for on-line merchandisers

4

## Recommendation

- It is not an new phenomenon arising from the digital era, but a **common social activity**.
- Customers tend to rely heavily on this service to reduce the amount of cognitive effort in making purchase decision.
- It provides online merchandisers new and powerful tools to influence customers' preferences and, ultimately, their purchase decisions.
- Higher personalized recommendation results in higher customer loyalty, higher sales and the benefit of targeted promotions.

5

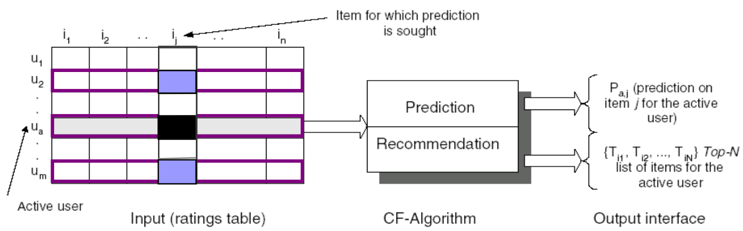
## Recommendation Approaches

- Popularity based
- Content based
- Collaborative filtering based
- Association based
- Demographics
- Reputation based or trust based

6

# Collaborative Filtering Recommendation

- The most successful and widely adopted recommendation approach.
- The principle of this approach is to **find users (consumers) with similar affinities** and rely on the preferences of these "neighbors" to provide recommendations.




General Process of Collaborative Filtering Approach [Sarwar et al., 2001]

7

# Research Motivation

- For the collaborative filtering approach, all ratings of items are considered identically important and given an equal weight in computing user similarities and identifying nearest neighbors for the active user.
- The collaborative filtering approach does not consider the **proximities between items (**i.e. item heterogeneity).
- However, we *believe* item heterogeneities should influence the recommendation effectiveness (prediction accuracy and coverage).
  - Movie (romance, horror, comedy, suspense, war, drama, etc.)
  - Text Books (data mining, data warehouse, internet marketing, electronic commerce, system analysis and design, technology management, decision support systems, etc.)

8

## Research Objective

- **Consider item heterogeneities in making item recommendations**

  The user preferences on items similar to the target item would be more reliable when predicting the user reference of the target item.
- We propose and develop a item **cluster-based collaborative filtering (CCF) recommendation approach** to recognize item heterogeneities.
- We integrate clustering techniques into collaborative filtering recommendation approach.

9

## Literature Review – CF approach

- In a typical collaborative filtering recommendation scenario, there is a set of $n$ **users** $U = \{u_1, u_2, \ldots, u_n\}$ and a set of $m$ **items** $I = \{i_1, i_2, \ldots, i_m\}$. Each user $u_i$ has a list of items $I_{u_i}$ (where $I_{u_i} \subseteq I$ and $I_{u_i}$ can be an empty set) on which the user has expressed his/her **preferences**.
- The preference of a user $u_i$ on item $i_j$ (donated as $p_{ij}$) can be explicit ratings (binary or numerical scale) provided by users or an implicit measure inferred from available user activities (i.e., purchase history, web logs, cookies, bookmarks, navigation patterns, and so on).

10

## Literature Review – CF approach (cont'd)

- The general process of a typical neighborhood-based collaborative filtering recommendation approach can be divided into three phases:

Dimension Reduction → Neighborhood Formation → Recommendation Generation

- Dimension Reduction
  - Lower Dimensional User Preference Matrix
- Neighborhood Formation
  - Computation of user similarity
  - Neighbor selection
- Recommendation Generation
  - Prediction scores
  - Top N list

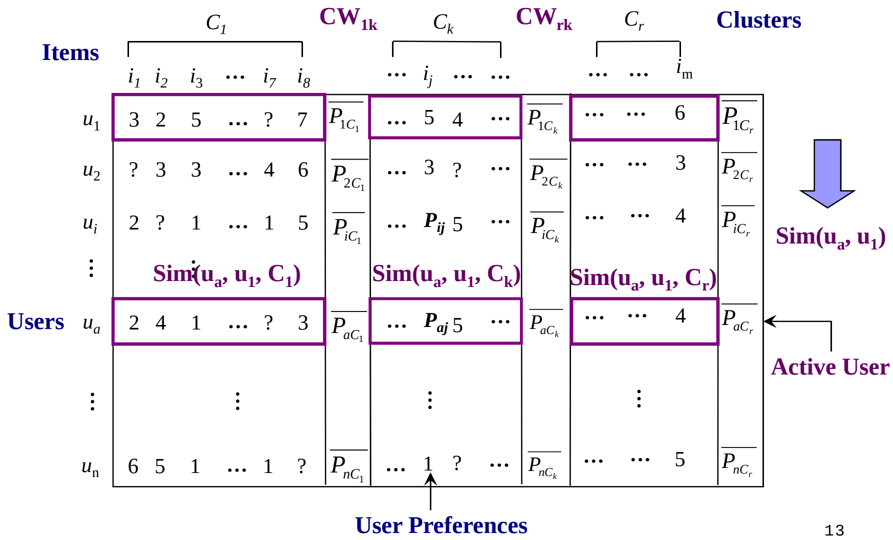
11

## Literature Review – Clustering

- Three main clustering approaches:
  - Partitioning-based (K-means, PAM, CLARA, etc.)
  - Hierarchical (HAC, HDC)
  - Neural-network-based (SOM)
- Document clustering
  - Text pre-processing is needed to transform each textual document into a feature vector first.

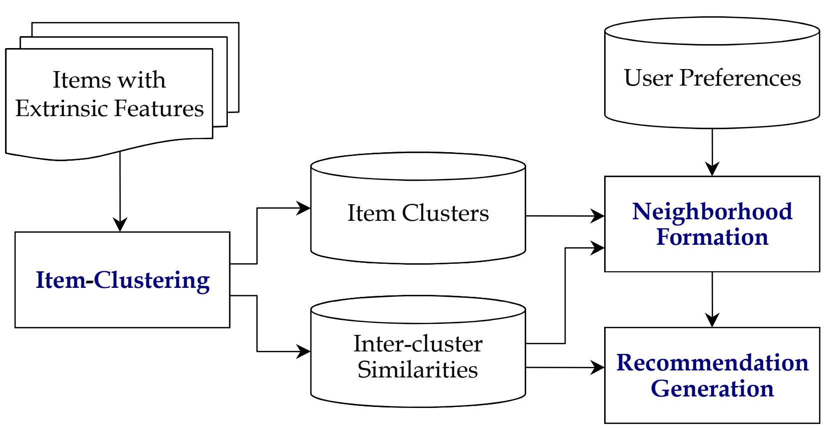
12

# Cluster-based Collaborative Filtering Recommendation Framework



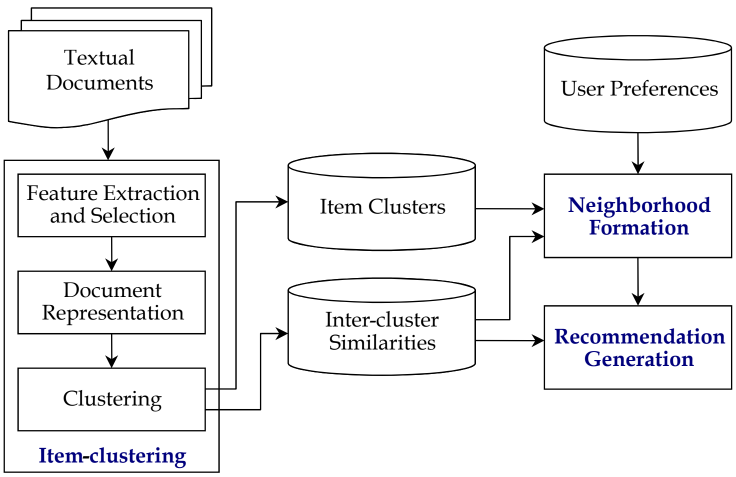

13

# Cluster-based Collaborative Filtering Approach for Recommending Items with Extrinsic Features




14

## Cluster-based Collaborative Filtering Approach for Recommending Textual Documents




15

# Item-clustering

- It groups items into **distinct clusters** and generates **inter-cluster similarities** for all pairs of clusters.
- A clustering algorithm can be applied directly to group items with extrinsic features. We employed the **hierarchical agglomerative clustering (HAC)** for item-clustering.
- Given any two clusters $C_r$ and $C_k$, their inter-cluster similarity (denoted as $CW_{rk}$) is estimated by the **group-average link method**. For every cluster $C_k$, $CW_{kk} = 1$.

16

# Neighborhood Formation

## Computation of User Similarity

- Three type of information for computing user similarity:
  - User preferences
  - Item clusters
  - Inter-cluster similarities
- For an active user $u_a$, instead of computing the similarity of the preference scores on co-rated items of $u_a$ and those of the other user $u_b$, several "**within-cluster" similarities** are computed first.

17

# Neighborhood Formation

## Computation of User Similarity (cont'd)

- Two similarity measures can be used to calculate the similarity of $u_a$ and $u_b$ within the cluster $C_r$:
  - Pearson correlation coefficient:

$$Sim(u_a, u_b, C_r) = \frac{\sum_{i \in C_r} \left(P_{ai} - \overline{P_{aC_r}}\right)\left(P_{bi} - \overline{P_{bC_r}}\right)}{\sqrt{\sum_{i \in C_r} \left(P_{ai} - \overline{P_{aC_r}}\right)^2} \sqrt{\sum_{i \in C_r} \left(P_{bi} - \overline{P_{bC_r}}\right)^2}}$$

  - Cosine:

$$Sim(u_a, u_b, C_r) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\left\|\vec{a}\right\|_2 \left\|\vec{b}\right\|_2} = \frac{\sum_{i \in C_r} p_{ai} \cdot p_{bi}}{\sqrt{\sum_{i \in C_r} p_{ai}^2} \sqrt{\sum_{i \in C_r} p_{bi}^2}}$$

18

## Neighborhood Formation

### Computation of User Similarity (cont'd)

- To predict the preference score of the active user $u_a$ on a target item that belongs to the cluster $C_k$, the overall similarity of $u_a$ and $u_b$ is estimated as:

$$Sim(u_a, u_b) = \frac{\sum_{r=1}^{c} Sim(u_a, u_b, C_r) \times CW_{rk}}{\sum_{r=1}^{c} CW_{rk}}$$

19

## Neighborhood Formation

### Neighborhood Selection

- Determine the best neighbors (similar users) of the active users.
- Neighbor selection methods:
  - Weight thresholding method uses an absolute threshold to find neighbors.
  - Center-based best-k method selects the k nearest users (k is pre-specified).

20

## Recommendation Generation

- Predict the preference of target item $i_j$ of active user $u_a$ according to the known preferences of the neighbors
- Adopt and modify the **deviation-from-mean method** where the mean is the cluster average rather than the overall average for such preference prediction
- The predicted preference of $u_a$ on a target item $i_j$ belonging to the cluster $C_k$ is defined as:

$$p_{aj} = \overline{p_{aC_k}} + \frac{\sum_{b=1}^{n}(p_{bj} - \overline{p_{bC_k}}) \cdot sim(u_a, u_b)}{\sum_{b=1}^{n} sim(u_a, u_b)}$$

21

## Experiment: Movie Dataset

- Items with extrinsic features:
  - 539 movie items (http://movie.kingnet.com.tw/)
  - 155 business undergraduate students participated
  - Each subject rated 50 randomly sampled movies with 7-point rating scale.
  - 5082 reliable ratings resulted from 103 reliable subjects
  - Sparsity level: 90.8 % (1-5082/(537*103))
  - extrinsic features:
    - 9 kinds of movie type (Category Attribute)
    - 4 kinds of movie rating; G, PG, PG-13 and R (Rank Attribute)
    - award record (Boolean)

22

# Experiment: Literature Dataset

- Textual documents:
  - 435 article items (chosen from DSS, ISR, JMIS, MISQ between 1999 and 2003)
  - 51 IS master/doctoral students participated
  - Subjects saw the title, abstract, authors, keywords and publication sources for rating.
  - 2244 reliable ratings resulted from 45 reliable subjects
  - Sparsity level: 88.5% (1-2244/(434*45))

23

# Evaluation Metrics of Recommendation Effectiveness

- *Accuracy:* **Mean Absolute Error (MAE)**
  - For each rating pair $< p_i, q_i >$, MAE uses the absolute error between them (i.e., $|p_i - q_i|$ ) and the MAE is calculated by summing up *N* absolute errors:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

- *Coverage*: Coverage is a measure of the percentage of items for which a recommendation system can provide predictions.
  - Because there may be no available neighbors for certain items, those items can not be predicted, resulting in lower prediction coverage.

24

## Empirical Evaluation Procedure

- For each user rating, we use ratings of all other subjects to predict this rating. The accuracy of predictions can be measured by MAE metric.
- Cluster each data set
- The best neighbors is determined by the *center-based best-k neighbors* method for each rating item.

25

## Cluster Data Sets for Evaluation Experiments

| Movie Dataset | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cluster Similarity Threshold** | **Original CF** | 0.52 | 0.56 | 0.57 | 0.58 | 0.59 | 0.60 | 0.61 | 0.66 | 0.73 | 0.80 |
| **Number of clusters** | 1 | 3 | 4 | 6 | 7 | 9 | 13 | 15 | 17 | 20 | 23 |

| Literature Dataset | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cluster Similarity Threshold** | **Original CF** | 0.02 | 0.04 | 0.06 | 0.08 | 0.10 | 0.12 | 0.14 | 0.16 | 0.20 | 0.26 |
| **Number of clusters** | 1 | 4 | 7 | 13 | 16 | 22 | 30 | 34 | 37 | 39 | 41 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Cluster Similarity Threshold** | 0.28 | 0.30 | 0.32 | 0.34 | 0.38 | 0.40 |
| **Number of clusters** | 47 | 56 | 66 | 70 | 81 | 95 |

26

## Evaluations

- Neighborhood Formation:
  - User Similarity Measure
  - Fill-in Strategy
- Effects on Recommendation Effectiveness
  - Neighbor size
  - Cluster size
  - Sparsity level

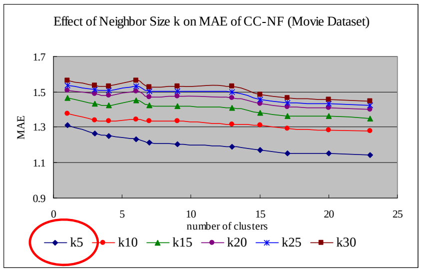
27

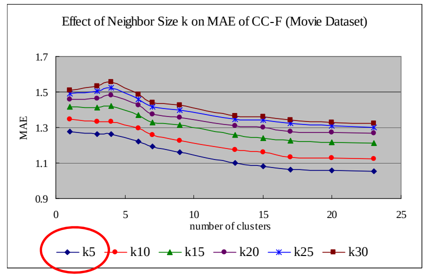
## Neighborhood Formation Methods

| Notation | Similarity Measure | Fill-in Strategy |
|---|---|---|
| CC-NF | Correlation Coefficient | No Fill |
| CC-F | Correlation Coefficient | Fill |
| CS-NF | Cosine | No Fill |
| CS-F | Cosine | Fill |

28

# Effect of Neighbor Size (CC-movie)

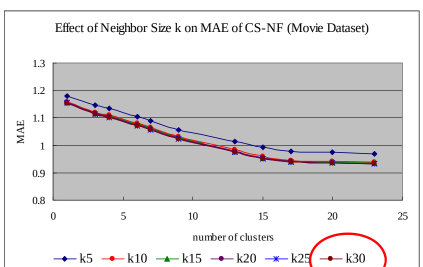


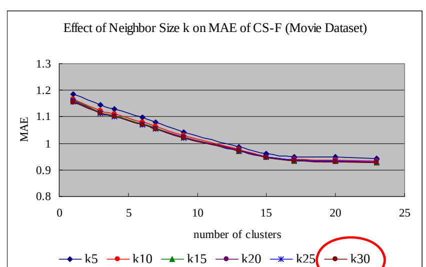




Thus, if using correlation coefficient as similarity measure, small neighbor size (e.g., 5) can lead to less MAE.

29

# Effect of Neighbor Size (CS-movie)




Thus, if using cosine as similarity measure, the larger neighbor size (e.g., 30) is better.

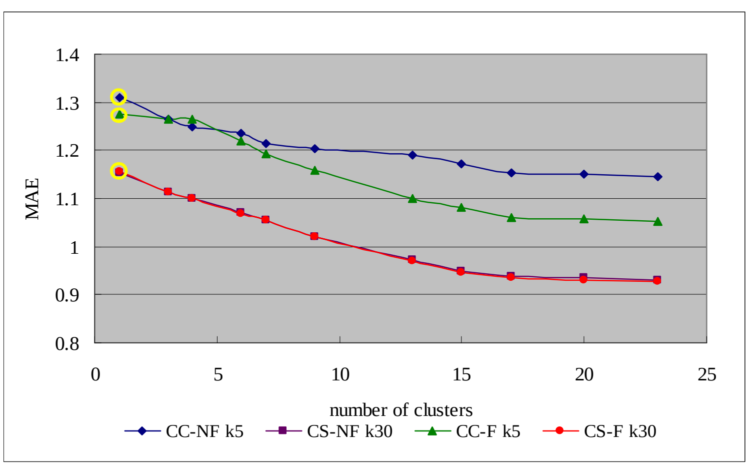
30

# Determined Neighbor size

| Neighbor Formation Method | Movie | Literature |
|---|---|---|
| CC-NF | 5 | 5 |
| CC-F | 5 | 5 |
| CS-NF | 30 | 20 |
| CS-F | 30 | 20 |

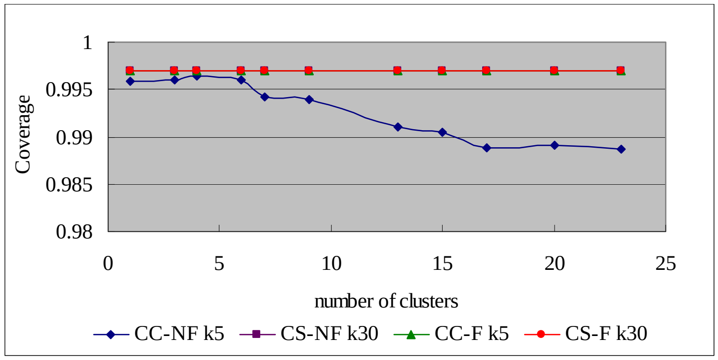
31

# Comparative Evaluation with Collaborative Filtering Approach on MAE (Movie)



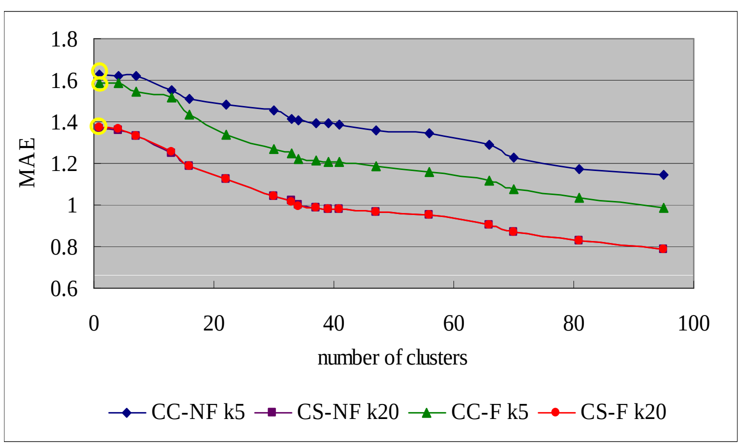

32

## Comparative Evaluation with Collaborative Filtering Approach on Coverage (Movie)



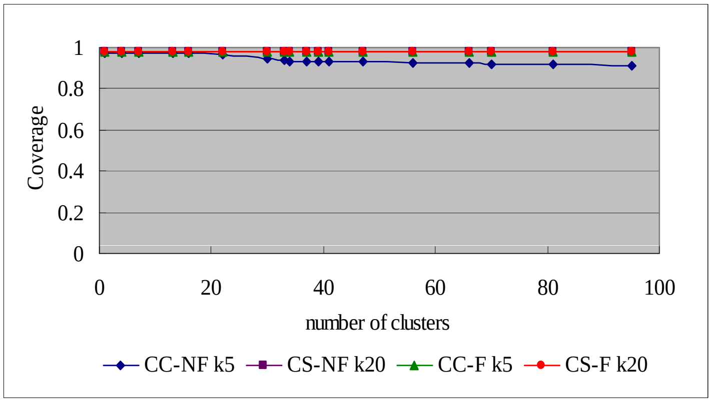

33

## Comparative Evaluation with Collaborative Filtering Approach on MAE (Literature)




34

# Comparative Evaluation with Collaborative Filtering Approach on Coverage (Literature)



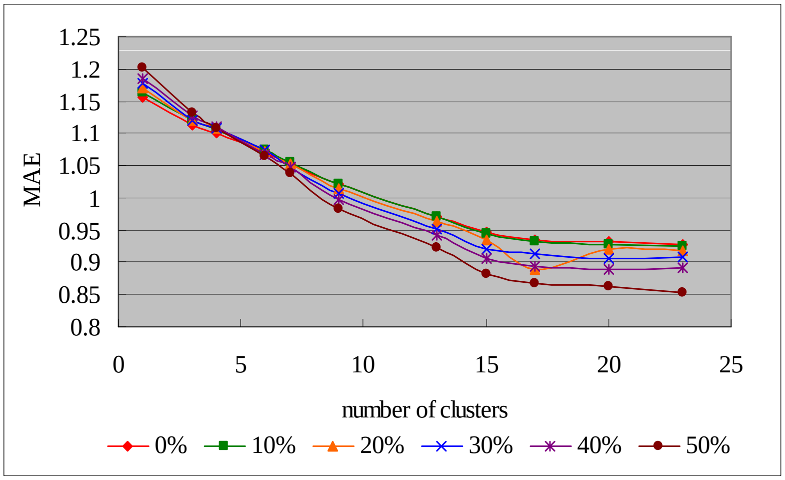

35

# Comparative Evaluation Summary

- The **cluster-based collaborative filtering approach** indeed improves prediction accuracy of collaborative filtering approach without sacrificing the prediction coverage.
- Using **cosine** similarity measure with fill-in achieved the best prediction performance in our experiments.
- The influence of **fill-in strategy** is more obvious when applying correlation coefficient similarity measure.

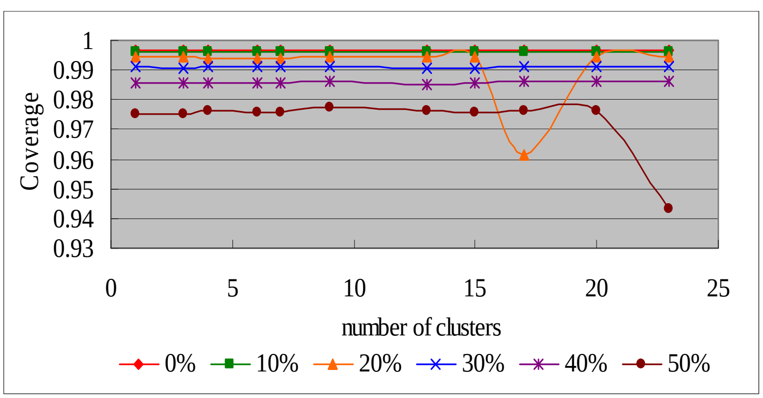
36

# Effect of Sparsity Level on MAE (Movie, CS-F)




37

# Effect of Sparsity Level on Coverage (Movie, CS-F)




38

## Sparsity Accommodation

The effect of **sparsity level** indicates that when the user preference is very sparse, the proposed cluster-based collaborative filtering approach with an ideal number of clusters can get much better prediction performance than the original collaborative filtering approach.

39

## Conclusions

- By considering the item heterogeneities, the **cluster-based collaborative filtering approach** indeed improves the prediction accuracy of collaborative filtering approach without sacrificing the prediction coverage.
- Moreover, in our experiments, adopting **cosine** similarity measure with **fill-in strategy** can achieve the best prediction performance.

40

## Future Research Directions

- Additional empirical evaluations in comparison with other techniques to improve the collaborative filtering recommendation approach in order to know the strengths and limitations of our cluster-based approach.
- Evaluate and apply our proposed cluster-based approach to real-world recommendations.
- Establish a rigorous foundation (may be a design-oriented theory) for item recommendation.

41

## Thank you

42