ClustKNN: A Highly Scalable Hybrid Model-& Memory-Based CF Algorithm

Al Mamunur Rashid, Shyong K. Lam, George Karypis, and John Riedl

University of Minnesota

GroupLens Research

Problem Domain

- Collaborative filtering (CF)-based recommender systems (RS).

- Issue:
  - Scalability

GroupLens Research

Background: Why Recommender Systems?

Information overload:

More than **1.3 million** articles!

About **50 million** blogs!

About **130 million** photos!

Background: Why Recommender Systems?

- One solution:
  - Recommender systems
    - Tools that suggest items of interest based on
      - Users' expressed preferences
      - Observed behaviors
      - Information about the items
    - **Collaborative Filtering**
      - Recommendations based on like-minded users

GroupLens Research

Many CF Algorithms So Far…

- Most of the early ones: kNN
  - GroupLens(1994), Ringo(1995)
- View it as a **special regression** problem.
  - Nearly all statistical and ML approaches can be applied!

- Classification by Breese et al.(1998):

	Memory-based CF	Model-based CF
Simplicity	👍	👎
Training cost	👍	👎
Online prediction cost	👎	👍
Adding new information	👍	👎

GroupLens Research

Many CF Algorithms So Far…

- Accuracy:
  - So far the main focus
    - However, how much difference in accuracy users perceive?

- Does it **scale** though?

GroupLens Research

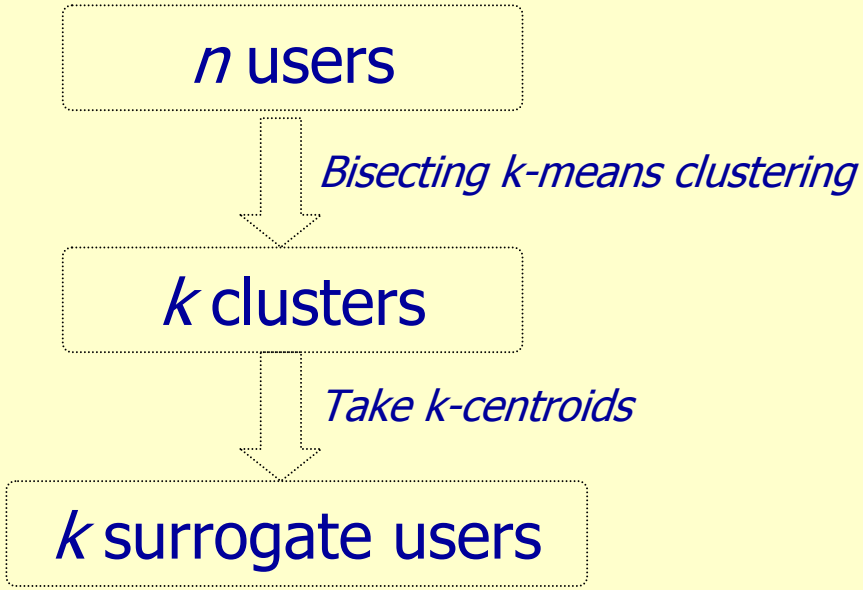
User-based *k*NN CF Algorithm

- Classic memory-based CF
- Assumption:
  - **Linear** relationship between two users' preferences
    - User-similarities measured by Pearson correlation coeff.
- Works very well
  - Very good **accuracy** & **Explainable** to general users.
- Problem: **Doesn't scale!**
  - ***O(mn)*** **online cost**

GroupLens Research

ClustKNN: Proposed Approach

- Retain good properties of User-based kNN
- Make it to scale



- Online cost: ***O(km)*** ≅ ***O(m)***
  - (k«m, k«n)

GroupLens Research

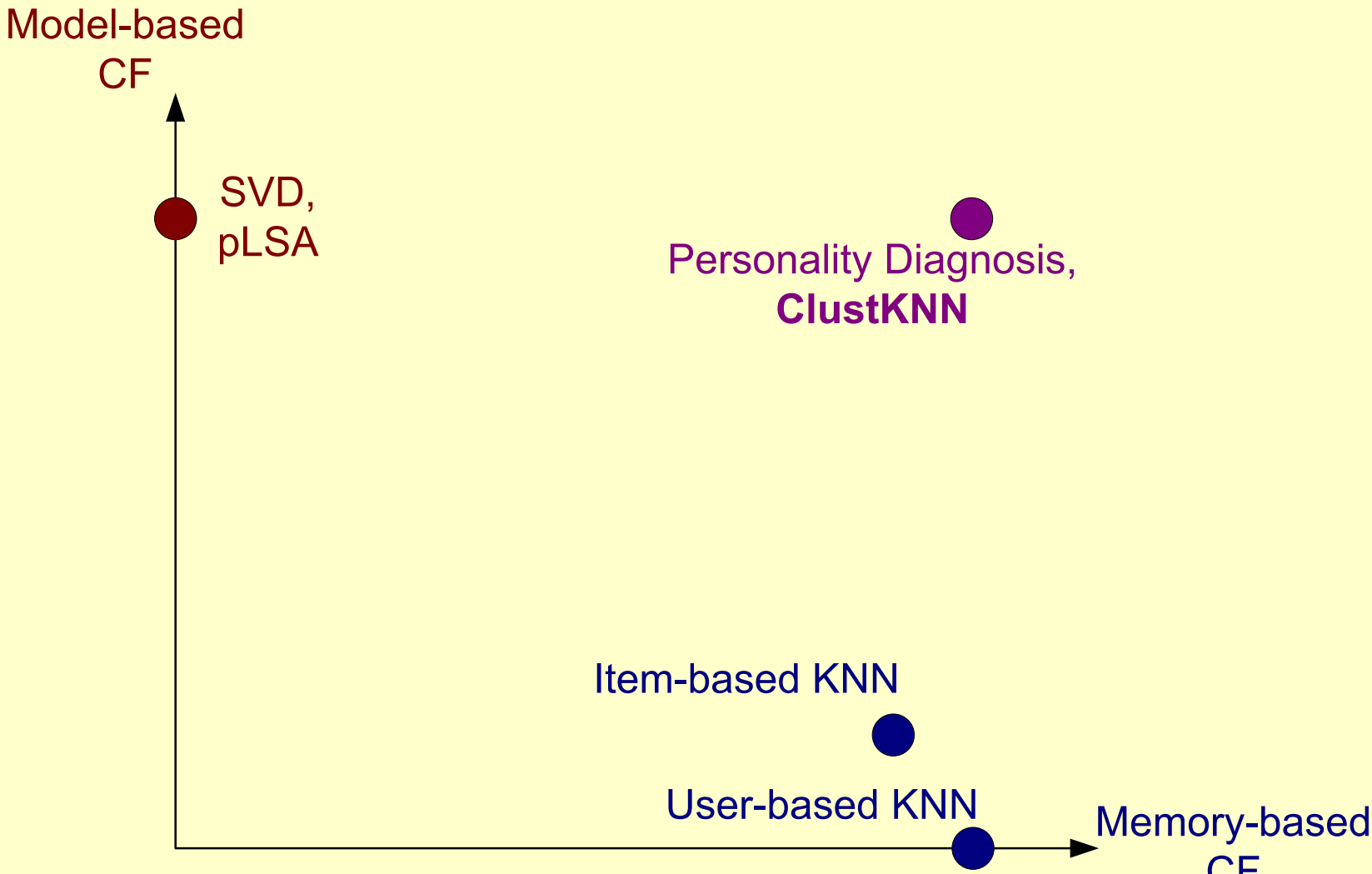
ClustKNN: Proposed Approach

- Bisecting k-means clustering
  - *Better* k-means
    - Cluster sizes are more uniform
    - Better results found in document clustering (Steinbach 2000)

- Similarity function:
  - Same in both cluster-building and CF
  - Nicely complements each other

GroupLens Research

Other Algorithms Considered



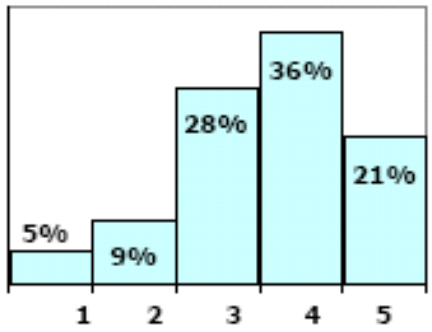
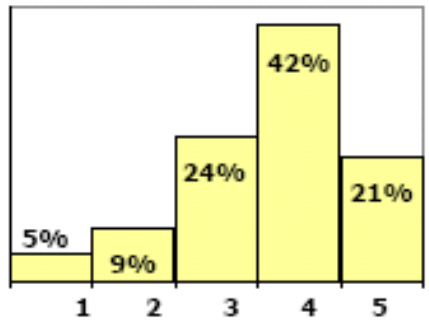
Time-complexities

CF algorithm	Offline	Online
pLSA	$O(mn)$	$O(m)$
SVD	$O(n^2m + m^2n)$	$O(m)$
Personality Diagnosis	-	$O(mn)$
ClustKnn	$\boldsymbol{O(mn)}$	$\boldsymbol{O(m)}$
User-based KNN	-	$O(mn)$
Item-based KNN	-	$O(mn)$

GroupLens Research

Experiments: Datasets

- Movie recommendation data from movielens (helping you find the *right* movies)

Property	ML1M	MLCURRENT
Number of users	6,040	21,526
Number of movies	3,706	8,848
Number of ratings	10,00,209	29,33,690
Minimum $\lvert u_i \rvert$, $\forall i$	20	15
Average rating	3.58	3.43
Sparsity	95.5%	98.5%
Rating distribution		

GroupLens Research

Experiments: Evaluation Metrics

- Prediction eval metrics
  - NMAE
    - Divide MAE with Expected MAE
    - Limitation:
      - Same value of error: same treatment
        - No difference between two (pred, actual) pairs (5, 2) and (2, 5)
  - Expected Utility (EU)

- Recommendation list eval metrics
  - Precision-recall-F1

GroupLens Research

Evaluation Metric: EU

- Two tables:
  - A contingency table
    - Rows: predictions; columns: actual ratings
  - A utility table
    - Filled with a linear utility function:

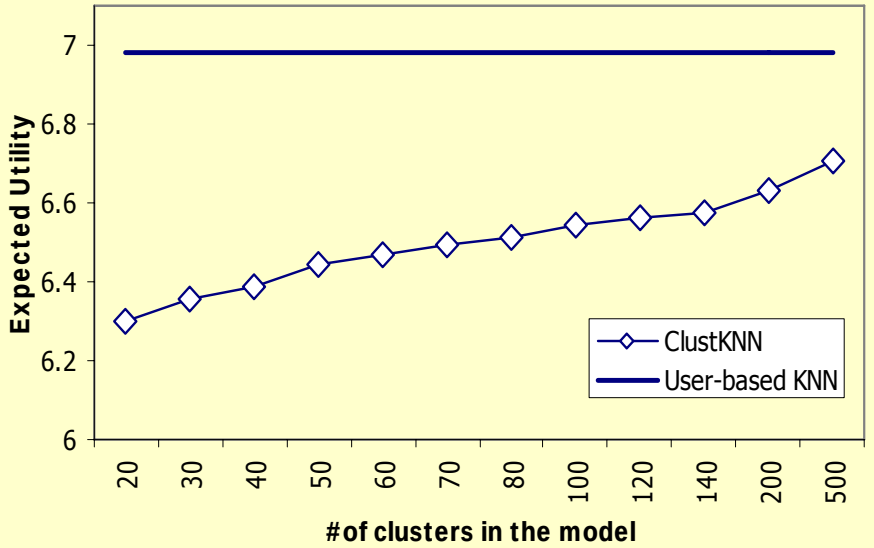
$$U(\hat{R}_i, R_j) = R_j - 2|\hat{R}_i - R_j|$$

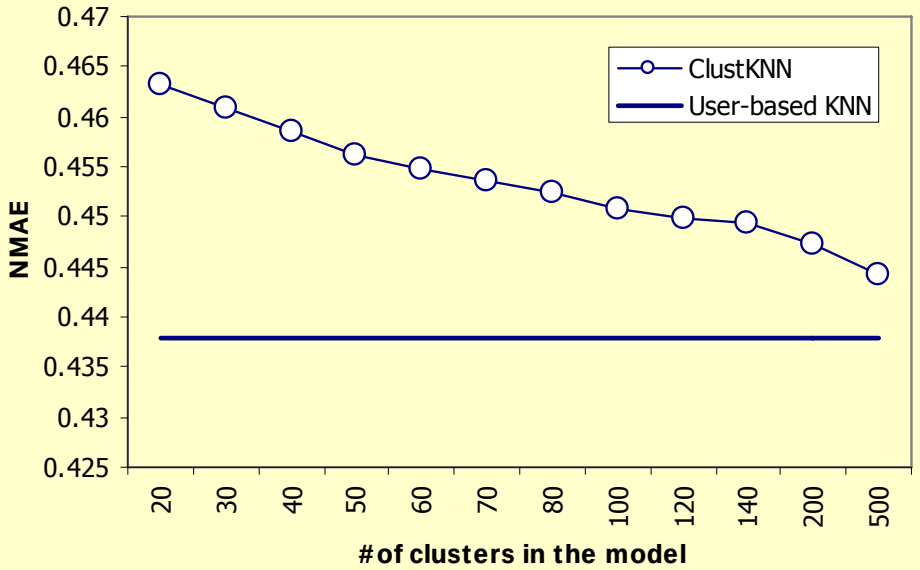
    - Penalizes false positives more than false negatives

$$EU = \sum_{\substack{1 \le i \le 10 \\ 1 \le j \le 10}} U(\hat{R}_i, R_j) P(\hat{R}_i | R_j)$$

GroupLens Research

Results





GroupLens Research

Results: Prediction Accuracy

CF algorithm	MAE		NMAE		EU	
	ML1M	MLCURRENT	ML1M	MLCURRENT	ML1M	MLCURRENT
SVD	**0.69**	-	**0.43**	-	6.81	-
User-based KNN	0.70	0.61	0.44	0.37	**6.98**	8.44
Item-based KNN	0.70	**0.60**	0.44	**0.36**	6.93	**8.48**
CLUSTKNN (k=200)	**0.72**	**0.62**	**0.45**	**0.37**	**6.63**	**7.82**
pLSA	0.72	0.61	0.45	0.37	6.57	7.95
Personality Diagnosis	0.77	0.66	0.48	0.40	5.00	3.19

GroupLens Research

Results: Recommendation List

CF algorithm	top-3				top-10			
	Precision		F1		Precision		F1	
	ML1M	MLCURRENT	ML1M	MLCURRENT	ML1M	MLCURRENT	ML1M	MLCURRENT
SVD	**0.8399**	-	**0.379**	-	**0.7564**	-	**0.6131**	-
User-based KNN	0.833	**0.6693**	0.379	**0.4086**	0.750	**0.5953**	0.610	0.556
Item-based KNN	0.819	0.657	0.374	0.407	0.749	0.592	0.610	0.556
CLUSTKNN (k=200)	**0.825**	**0.659**	**0.377**	**0.407**	**0.743**	**0.589**	**0.606**	**0.553**
pLSA	0.817	0.656	0.375	0.406	0.739	0.587	0.604	0.552
Personality Diagnosis	0.789	0.622	0.366	0.391	0.723	0.565	0.595	0.537

GroupLens Research

ClustKNN: Discussion

- Scalable!
- Simple and explainable
- Hybrid of model- and memory-based approaches
- Great for occasionally-connected, low-storage devices!
  - Memory requirement: only ***O(km+m)*** !

GroupLens Research

Thanks for listening!

Questions?

GroupLens Research