Implementation and Evaluation of Collaborative Filtering Algorithms

Group 1

Task 1	Model-based vs. Memory-based Algorithms
Task 2	Correlation/Similarity Measures
Task 3	Significance and Variance Weighting

1 Collective Filtering Algorithms

Model-based and Memory-based

Model-based vs. Memory-based

1. 10-fold cross-validation for best number of clusters from 2 to 12
2. Use the ranking score as the evaluation metric
3. Apply the best number of cluster in test data
4. Acquired ranking score of 39.35477, with whole time cost more than 2 hours

1. For each sample (user, item, etc.), calculate its similarity (definitions vary) to all other samples. Store all calculations in a square matrix.
2. When predicting, for each unknown attribute of each sample, fill in the sum of sample average attribute and neighbors' same attribute's deviation from their averages, weighted by their similarities.

Evaluation

- Movie dataset: Mean Absolute Error (MAE) as suggested in class
- Microsoft Dataset: Rank score as Expected Utility:
  - Utility as excess predicted preference (from 0 to 1-d) over the probability it is visited (modeled as exponential decay)
  - To discount contribution by training data, only users and attribute IDs present in the testing dataset are used, so values tend to be smaller than usual calculations

Result

		MAE (Movie data)	Rank Score (MS data)
Model-based	Performance		39.35 (different calculation method)
	Efficiency		2 hours for 10 fold cross validation
Memory-based (Pearson Correlation)	Performance	1.09	26.89
	Efficiency	Similarity 2hrs; Prediction 1hrs	Similarity 15min; Prediction ~15min

2 Correlation/Similarity Measures

Pearson
Vector Similarity
Entropy

Spearman
Mean Square Difference
SimRank

Spearman

- Evaluating monotonic relationships involving ordinal variables (ranks, ratings)
- Reasonable to consider because it doesn’t assume distance measurement model
- Paper suggests no significant improvement however

Vector Similarity (Cosine Similarity)

- Measures the angle between two vectors to calculate how “similar” they are
- Often used in text analysis

Entropy

- “Measure of association ... reduction in entropy … from knowing another user’s ratings” (p. 233)
- Measured entropy(A) - conditional entropy(A | B)
- The bigger the measure is, the more correlated A and B are

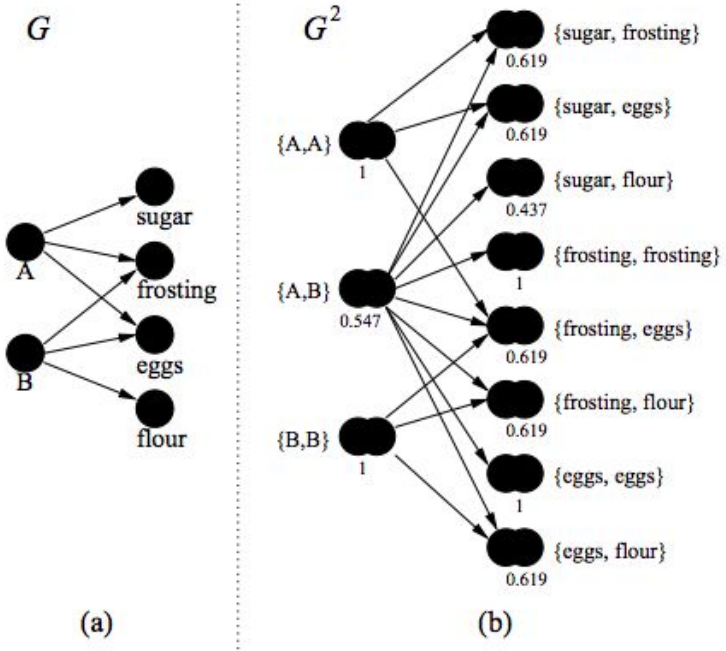
Mean Square Difference

- mean((rowA[joint_values] - rowB[joint_values])^2)
- Measure the difference between two rows
- Large value means small similarity
- Take inverse of the value

SimRank

- Measure of structural-context similarity
- Similarity based on the relationships with other objects
- “Two objects are similar if they are related to similar objects”
- S(A, B) as a function of s(items) (and vice versa)

Bipartite SimRank



|O(B)|

|O(A)|		
	0.619	0.619	0.437
	1	0.619	0.619
	0.619	1	0.619

Result

	MAE (Movie data)	Rank Score (MS data)
Pearson	1.09	26.89
Spearman	1.09	26.89
Vector Similarity	1.10	27.03
Entropy	1.10	26.91
Mean Square Difference	1.10	27.14
SimRank	1.08	27.08

Result

- Some minor improvements but overall not significant
- Some measures (e.g. entropy) a lot less efficient (about 10x calculation time)

3 Significance/Variance Weighting

Significance Weighting

- Take into account the amount of trust
- More common items = more trustworthy correlation
- Fewer than 50 commonly rated items → n/50 weight (1 otherwise)

Variance Weighting

- “If everyone likes Avatar, liking Avatar is not a great measure of correlation”
- *Item-wise* variance weight:
  (var - min(var)) / max(var) ($0 \leq w \leq 1$)
- The more varied the ratings for the item are, the more influential

Result

	MAE (Movie data)	Rank Score (MS data)
Pearson	1.09	26.89
Pearson w/ Significance Weighting	1.10	26.89
Pearson w/ Variance Weighting	1	26.3

Result

- No significant improvement with significance/variance weighting
- Variance Weighting does comparatively well on movie dataset
  - Differing taste for movies?
- Number of co-rated items → good indicator of similarity?
  could indicate similar taste but ratings could vary
- Spearman correlation may yield better result