Hybrid Recommendation System for Social News Websites using Collaborative and Content-Based Filtering

Viswanath S (106108011)
Sanjana Singh (106108079)
Srinivasan Raghavan

Recommendation System for Social News Websites

- Learns from user history (links clicked, upvoted, downvoted)
- Predicts new categories of links to users

Progress

- Implemented LSA-SVD technique for recommendation
- Experiments on datasets from www.reddit.com
- Working on ClustKNN - a hybrid model and memory based collaborative filtering method

Latent Semantic Analysis using Singular Vector Decomposition

(http://github.com/viswanathgs/reddit-recommender)

LSA-SVD Algorithm - Input

- Input: Document-Term Matrix, M (m x n)
- Rows: Users (m)
- Columns: Features (n)
- Features: Categories of links (eg., technology, movies)
- M[i][j] = affinity of user i towards attribute j
- Affinity: ratio of up-votes to total votes on links belonging to that category (normalized to [-1.0, 1.0])
- Positive affinity implies more up-votes than down-votes on that category

LSA-SVD Algorithm - Training

- Apply SVD on matrix M to factorize M into matrices U, S and V such that

$$M = U \times S \times V^T$$

where U (m x d), V (n x d), S (d x d)

$$d = \min(m, n)$$

- U: left singular matrix
- V: right singular matrix
- S: diagonal matrix of singular values

LSA-SVD Algorithm - Training

- Each row of U (d-columns) corresponds to the reduced dimensions of each user
- Each row of V corresponds to the reduced dimensions of each attribute (category)
- The rows of U and V are implicitly plotted in a d-dimensional space
- The users and categories that are close to each other can be clustered together

LSA-SVD Algorithm - Testing

- For each test-input (user, category), calculate the distance between the user and the category from the d-dimensional implicit graph
- Distance-metric: Euclidean distance (extended to d-dimensions)
- If distance < threshold, then prediction = "Yes", else prediction = "No"
- Optimize threshold to achieve maximum accuracy (using 10-fold cross-validation)

Experimental Dataset

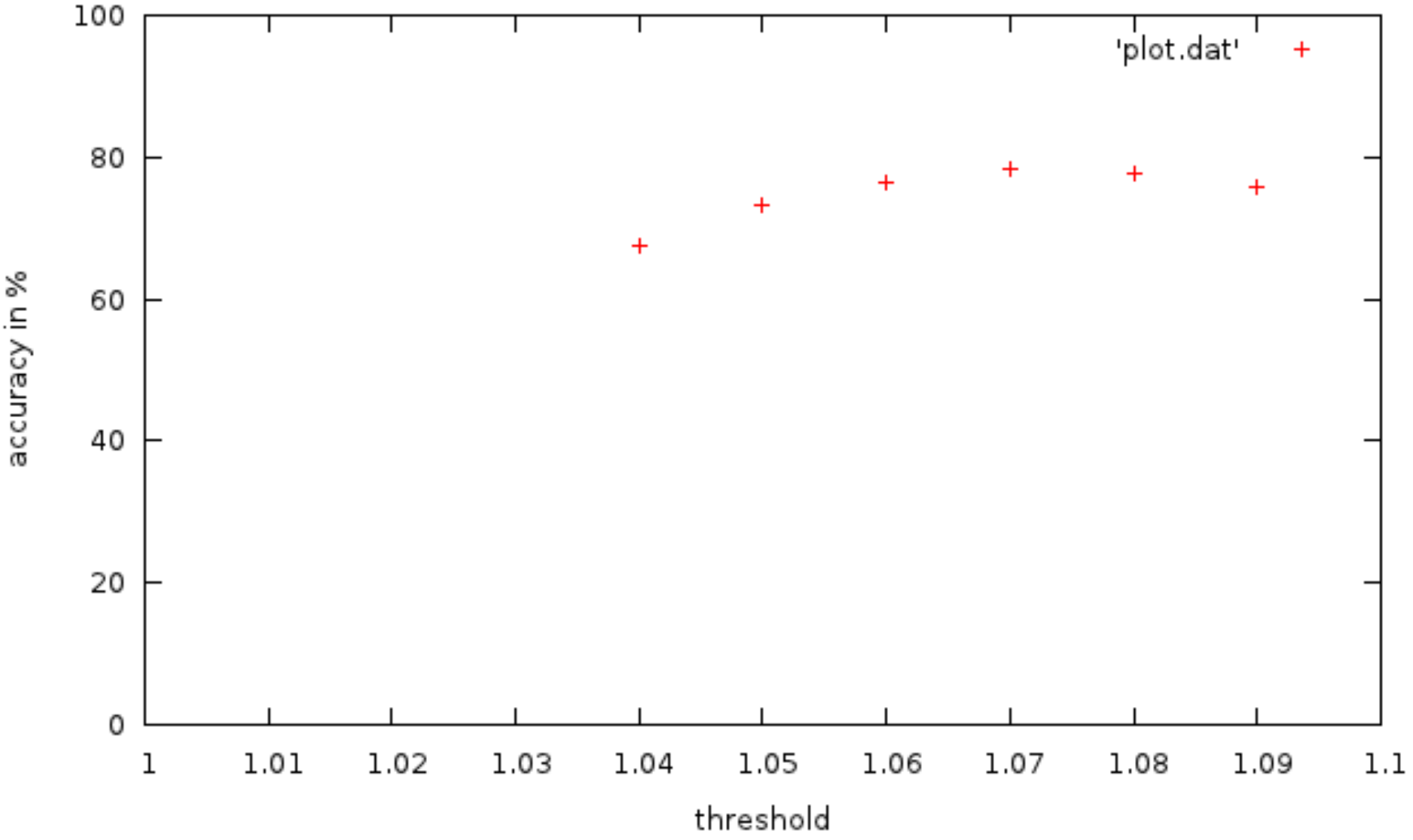
- Datasets obtained from www.reddit.com
- Processed the raw data to extract the user-id and the sub-reddit (category)
- Normalized the affinities to the range [-1.0, 1.0]
- Chose the most frequent 100 attributes (categories)
- Matrix of size 9160 x 100 (9160 users and 100 attributes)

Experimental Results

Accuracies obtained for different threshold values after 10-fold cross validation

Threshold	Accuracy
1.04	**67.45 %**
1.05	**73.15 %**
1.06	**76.50 %**
1.07	**78.26 %**
1.08	**77.62 %**
1.09	**75.65 %**

Experimental Results



ClustKNN: A Highly Scalable Hybrid Model and Memory Based Collaborative Filtering Algorithm, by Al Mamunur Rashid , Shyong K. Lam , George Karypis , John Riedl at 12th ACM SIGKDD

ClustKNN

- Memory-based algorithms (eg., KNN)
  - Utilize the entire dataset for each recommendation
  - Simple to implement and requires no training
  - Easy to accommodate changes in dataset
  - Disadvantage: Very slow for large datasets
- Model-based algorithms (eg., SVD)
  - Computes a model of the training dataset
  - The model is compact compared to the dataset
  - Faster recommendations based on the model
  - Disadvantage: Recompute model for changes in dataset
- ClustKNN: A hybrid of memory and model-based techniques

ClustKNN

- Training (Model-building)
  - Use bisecting k-means clustering (an improved version of basic k-means)
  - Compute the k surrogate users corresponding to the centroids of each cluster
- Testing (Prediction score calculation)
  - Compute similarity of target user with each centroid (using Pearson correlation coefficient)
  - Pick the top m centroids (surrogate users)
  - Compute prediction score from the selected centroids (using adjusted weighted average)

Thank You