11741 Reading Summary, Collaborative Filtering

Xin Qian (xinq@cs.cmu.edu)

CF assumes that given users X and Y who rate n items in a similar way, they tend to rate other items similarly. It predict user's preference towards new topics or products given past item preference history from existed users. Challenges in CF includes sparse data handling, prompt response and synonymy, shilling attacks, data noise and privacy protection problems. Memory-based CF suffers from the reliability issue when data are sparse. Model-based CF approach us the rating data to learn a model to make better prediction. Content-based filtering analyzes the textual content and explores the regularities in the content but ignores the preference similarity across individuals. Metrics used to evaluate CF includes MAE, Precision, recall and ROC sensitivity.

Data sparsity challenge arises on cold start problem where new items cannot be recommended and it is hard to recommend items to new users. Dimensionality reduction techniques such as Singular Value Decomposition removes users to reduce the dimensionalities. Eigentaste applies PCA to reduce dimensionality. However, this way of dealing with sparse data may degrade recommendation performance. Hybrid CF utilizes external content information to predict for new users or new items.

Scalability issues is not a problem for dimensionality reduction techniques, e.g. incremental SVD. Memory-based CF can achieve satisfactory scalability. Instead of computing all-pair similarity, item-based Pearson CF compute only the pair similarity between co-rated items by a user.

Synonym can be solved by the SVD techniques where we apply SVD on a large term-document association matrix. LSI partially resolves the polysemy problem. A per-user basis weighting scheme used in a hybrid approach combining content-based and CF recommendations helps to solve the grey sheep problem. Item-based CF algorithm was less affected by the attacks than the user-based CF algorithm. Personal privacy protection, increased noise and explainability are three challenges in recommender systems.

Similarity computation between items are done by finding users who have rated both of these items. User-based CF algorithm calculates similarity between two users who have both rated the same items. Correlation-based similarity computation is done by computing the Pearson correlation, which measures the extent to which two variables linearly relate with each other. Adjusted cosine similarity takes into account the fact that different users may use different rating scales. Weighted sum of others' ratings on the same item is used to predict for an active user on a certain item.

Bayesian Belief Net CF algorithms is a directed, acyclic graph with triplets. It is often used for classification tasks. A simple Bayesian CF can have worse predictive accuracy but better scalability than a Pearson correlation-based CF. Tree augmented Naive Bayes and Naive Bayes optimized by ELR received high classification accuracy for both complete and incomplete data. Bayesian belief nets with decision trees at each node has each node corresponds to each item in the domain and the states of each node correspond to the possible ratings for each item. Clustering CF models use clustering as an intermediate step before applying other memory-based CF algorithm. Sparse factor analysis replaces missing elements with default voting values to remedy the sparseness in user-item matrix. Viewing the recommendation process as a sequential optimization problem which uses a Markov decision process model are MDP-based CF algorithms.