Proposed Algorithm

1. Dataset.

The experiment has the Amazon review Dataset {McAuley}. Due to the large dataset size, only two categories are used to demonstrate the experiment. Digital Music and Health Care are the two categories used for experiment.

1. Accuracy Improvements.

The motive of the research is to find a better way to generate recommendations. These recommendations will be Top-N and the best prediction accuracy improvements.

1. Algorithm for kNN/SVD and the text review like LDA (Latent Dirichlet Allocation).

The k nearest neighbor / kNN classification determines the decision boundary for sample locally. Each samples is considered in the k classes of nearest neighbors.

TRAIN-KNN(C, D)

1 D’ 🡨 Preprocess(D)

2 k 🡨 Select-K(C,D’)

3 return D’, k

APPLY-KNN(C,D’,k,d)

1 Sk 🡨 ComputeNearestNeighbors(D’,k,d)

2 for each cj epsilon C

3 do p\_\_j 🡨 ||

Latent Semantic Indexing is a process of using SVD for approximating the lower rank representation which then enables to compute the query-document similarity scores in this low-rank representation.

In LSI, we use SVD to construct a low-rank approximation matrix C, to the term-document matrix, for a value of k that is far smaller than original rank of C.

M = U ∑ V\*

U – m \* m unitary matrix

∑ - diagonal m \* n matrix

V – n \* n unitary matrix, V\* is the conjugate transpose of V.

Aspect based Model for the use of reviews with the combination of ratings to improve recommendations. LDA for the term-frequency and the inverse-document frequency is used to make the reviews work in combination of the ratings per document(review).

1. Comparison to show the results.

Sources:

* <https://en.wikipedia.org/wiki/Singular-value_decomposition#Low-rank_matrix_approximation>
* <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>
* <http://www.aclweb.org/anthology/E17-2107>