ABSTRACT

DIMSUM, an efficient and accurate all-pair similarity algorithm for real-world large scale dataset, tackles shuffle size problem of several similarity measures using MapReduce. The algorithm uses a sampling technique to reduce 'power items' and preserves similarities. This work presents an improved algorithm DIMSUM+ with a complex sampling technique to enhance DIMSUM so that it is able to further reduce 'power users'. The algorithm generates *k*-nearest-neighbor matrix that is used in collaborative based Recommender systems. The evaluations of algorithm on MovieLens dataset with 1 million movie ratings and Yahoo! Music dataset with 700 million song ratings show significant improvement that DIMSUM+ outperforms DIMSUM at least 1.4x faster.