11741 Reading Summary, Ch 18

Xin Qian (xinq@cs.cmu.edu)

Linear algebra review

The rank of a matrix is the minumum number of linearly independent rows or columns. We could get the eigenvalues of a matrix by solving the characteristic equation. The effect of multiplication for an arbitrary vector by S is determined by the eigenvalues and eigenvectors of S. This section ends with the intuition interpretation on the effect of small eigenvalues and their eigenvectors on a matrix-vector product is small. A square matrix can be factored into the product of matrices derived from its eigenvectors. The low-rank approximations to term-document matrics are built on the symmetric diagonal decomposition theorem.

Term-document matrics

Term-document matrix C are likely to have M != N. And C is very likely to be asymmetric. Performing matrix decomposition on C multiplied by C transpose (a square symmetric matrix) gives the singular value decomposition, which is an extension of the symmetric diagonal decomposition.

Low-rank approximations

This section states a matrix approximation problem where we want to find a M by N matrix of rank at most k to minimize the Frobenius norm (discrepancy) of the matrix difference. Replacing small eigenvalues by zero creates the best rank-k approximation to C. This section ends with discussing why truncating the smallest r-k singular values helps generate low error rank-k approximation.

Latent semantic indexing

LSI sometimes is refered to LSA as well. Casting queries into the low-rank representation and computing corresponding query-document similarity scores is known as latent semantic indexing. Recall that the vector space representation sufferes from the inexpressiveness of synonymy and polysemy as well as the large size of the term-document matrix. We might want to use the SVD to construct a low-rank approcimation matrix to the original term-document matrix. A popular choice of k is usually in the low hundreds. LSI for query-document similarity matches was as follows, we map a query in the LSI space by the SVD transformation on the term-document matrix. This process is not specific to only query vector, but can be a document vector who wishes to be added in the collection. We are expecting to retain the retrieval quality from dimension reduction and improve such quality by grouping together similar co-occurrence terms. LSI has two major drawbacks of vector space retrieval: unable to express negations and unable to enforce Boolean conditions. Experiment results expects and verified that reducing k tends to increase recall. LSI can be interpreted as a soft clustering and each document has a fractional membership in each dimension.