Abstract

With large sets of text documents increasing rapidly, being able to efficiently utilize this vast volume of new information and service resource presents challenges to computational scientists. Text documents are usually modeled as a term-document matrix which has high dimensional and space vectors. To reduce the high dimensions, one of the various dimensionality reduction methods, concept decomposition, has been developed by some researchers. This method is based on document clustering techniques and leastsquare matrix approximation to approximate the matrix of vectors. However the numerical computation is expensive, as an inverse of a dense matrix formed by the concept vector matrix is required. In this paper we presented a class of multistep spare matrix strategies for concept decomposition matrix approximation. In this approach, a series of simple sparse matrices are used to approximate the decompositions. Our numerical experiments on both small and large datasets show the advantage of such an approach in terms of storage costs and query time compared with the least-squares based approach while maintaining comparable retrieval quality.
Multistep Sparse Approximation Technology in Information Retrieval

References

- M. Berry and M. Browne. Understanding Search Engines: Mathematical modeling and text retrieval. SIAM, 1999
Multistep Sparse Approximation Technology in Information Retrieval


Index Terms

Computer Science

Information Sciences

Keywords

Term-document matrix  Concept decomposition matrix  Sparsity pattern  Multistep Sparse matrix approximation