

**CHENNAI – PONDICHERRY**

**A Holistic Approach to Distributed Dimensionality Reduction of Big Data**

**Abstract:**

With the exponential growth of data volume, big data have placed an unprecedented burden on current computing infrastructure. Dimensionality reduction of big data attracts a great deal of attention in recent years as an efficient method to extract the core data which is smaller to store and faster to process. This paper aims at addressing the three fundamental problems closely related to distributed dimensionality reduction of big data, i.e. big data fusion, dimensionality reduction algorithm and construction of distributed computing platform. A chunk tensor method is presented to fuse the unstructured, semi-structured and structured data as a unified model in which all characteristics of the heterogeneous data are appropriately arranged along the tensor orders. A Lanczos based High Order Singular Value Decomposition algorithm is proposed to reduce dimensionality of the unified model. Theoretical analyses of the algorithm are provided in terms of storage scheme, convergence property and computation cost. To execute the dimensionality reduction task, this paper employs the Transparent Computing paradigm to construct a distributed computing platform as well as utilizes the linear predictive model to partition the data blocks. Experimental results demonstrate that the proposed holistic approach is efficient for distributed dimensionality reduction of big data.

**Existing System:**

PCA applies the linear operations to transform the data to a new coordinate system in which the greatest variance comes to lie on the first coordinate, the second on the next coordinate, and so on. Singular Value Decomposition (SVD) is a method to identify the dimensions along which data points exhibit the most variation. SVD allows to find the best approximation of the original data using fewer dimensions. Independent Component Analysis (ICA) is a high-order method that seeks linear projection, not necessarily orthogonal to each other, that are as nearly statistically independent as possible. Factor analysis is a statistical method to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.

**Proposed System:**

We employ the Transparent Computing paradigm to construct a distributed computing platform with heterogeneous devices to perform the dimensionality reduction tasks of big data. A linear predictive model is used to estimate the size of data blocks that will be distributed to autonomic devices.

We present a Lanczos based High Order Singular Value Decomposition (L-HOSVD) algorithm that can quickly obtain the core tensor and the truncated orthogonal bases from the unified tensor model. The tensor unfolding are transformed to symmetric matrices and the Lanczos method is employed to compute the singular vectors and singular values. According to our experiment, the L-HOSVD algorithm is efficient and competitive.