Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing Letters 94 (2005) 247–252

Information Processing Letters

www.elsevier.com/locate/ipl

Exact k-NN queries on clustered SVD datasets

Alexander Thomasian [*,1], Yue Li, Lijuan Zhang

Computer Science Department, New Jersey Institute of Technology (NJIT), Newark, NJ 07102, USA

Received 10 December 2004; received in revised form 7 March 2005

Communicated by J. Chomicki

Abstract

Clustered SVD–CSVD, which combines clustering and singular value decomposition (SVD), outperforms SVD applied globally, without first applying clustering. Datasets of feature vectors in various application domains exhibit local correlations, which allow CSVD to attain a higher dimensionality reduction than SVD for the same normalized mean square error. We specify an exact method for processing k-nearest-neighbor queries for CSVD, which ensures 100% recall and is experimentally shown to require less CPU processing time than the approximate method originally specified for CSVD.

Keywords: Singular value decomposition; Principal component analysis; Dimensionality reduction; Clustering; Multidimensional indexing; k-nearest-neighbor queries; Algorithms; Databases

1. Introduction

Content based retrieval (CBR) is concerned with retrieving objects represented by their feature vectors. Content based image retrieval (CBIR) is a special case, where features are based on texture, color, shape, etc. [1]. Similarity of two objects is determined by the proximity of the endpoints of the N-dimensional feature vectors representing them. Proximity is determined by the Euclidean distance or some other similarity measures [2].

Range and k-nearest-neighbor (k-NN) queries are two popular similarity search paradigms. Range queries retrieve all points within distance ε of a query point, while k-NN queries retrieve the k objects with the closest feature vectors to the query point. It is easy to see that a k-NN query is tantamount to the processing of a range query with an appropriate radius, but a k-NN query has the advantage that there is no need to specify ε.

For a very large number of objects and a high dimensionality for feature vectors, the processing of k-NN queries on an $M \times N$ dataset via a sequential scan can be quite costly. CPU time is dominated by

[*] Corresponding author.
E-mail addresses: athomas@cs.njit.edu (A. Thomasian), yue@cs.njit.edu (Y. Li), lijuan@cs.njit.edu (L. Zhang).

[1] We acknowledge the support of NSF through Grant 0105485 in Computer System Architecture.

0020-0190/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ipl.2005.03.003

the cost of computing Euclidean distances, since the cost of inserting nearest neighbor candidates into a heap is negligibly small. It is especially important to minimize the number of objects inspected for more complex similarity measures [2].

Singular Value Decomposition (SVD) method or *Principal Component Analysis* (PCA) are equivalent methods, which can be used for dimensionality reduction [5]. The processing of k-NN queries with a reduced number of dimensions is less costly, but yields approximate results, as quantified by recall and precision [5] (see Section 2). Most multidimensional indexing structures [6] lose their efficiency when processing k-NN queries on high dimensional data [5,2], so dimensionality reduction is another method to cope with the "curse of dimensionality". This paper is concerned with the cost of reducing CPU time using linear scans of associated datasets. Multidimensional indexing methods to reduce query processing cost are discussed in Section 4.

Given an $M \times N$ matrix X of feature vectors, PCA first computes and then decomposes its covariance matrix: $C = X^t X/M = V \Lambda V^t$, where V holds the eigenvectors, and Λ is the diagonal matrix of eigenvalues [5,7]. The columns of X have a zero mean and studentization (division by the standard deviation) before applying SVD or PCA is required when the columns of X have highly variable magnitudes. All eigenvalues are positive, since the covariance matrix is positive semi-definite [7]. We assume, without loss of generality, that the eigenvalues λ_i, $1 \leqslant i \leqslant N$, are in non-increasing order. The rank of the matrix is determined by the number of eigenvalues which are not close to zero.

Alternatively, the singular value decomposition for matrix X is: $X = USV^t$, where S is the diagonal matrix of singular values: $s_n = \sqrt{M\lambda_n},\ 1 \leqslant n \leqslant N$, and V is the matrix of eigenvectors as before. The *Normalized Mean Square Error* (NMSE) is computed with respect to $Y = XV$, given that $p \leqslant N$ of its features (columns) are retained:

$$\mathrm{NMSE} = \frac{\sum_{n=p+1}^{N} \lambda_n}{\sum_{n=1}^{N} \lambda_n} = \frac{\sum_{m=1}^{M} \sum_{n=p+1}^{N} y_{m,n}^2}{\sum_{m=1}^{M} \sum_{n=1}^{N} y_{m,n}^2}.$$

The Karhunen–Loeve transform, which minimizes the NMSE for a given number of retained dimensions is only applicable to static datasets [5]. A straightforward application of SVD or PCA to the original dataset benefits from global correlations, but most real-life datasets exhibit local correlations, which benefit the *Clustering and Singular Value Decomposition* (CSVD) method [3]. In fact clustering and SVD can be applied recursively, even starting with SVD [14,3]. Experimental results show that for the same value of the NMSE, CSVD achieves a higher dimensionality reduction, i.e., a fewer number of retained dimensions than when SVD alone is applied to the global dataset. Conversely, given the number of dimensions to be retained p overall clusters, the minimum NMSE is attained [15,3].

A survey of methods related to CSVD is given in [3], but the method in [8], which deals with the dynamic insertion of data is also very relevant. Applying SVD to dynamic data is unacceptably expensive, since the SVD computation would have to be repeated for each insertion. Clustering the dataset, e.g., by building a multidimensional index, allows this computation to be done on a subset of data. To save costs even further points can be initially inserted into the appropriate clusters, or nodes of the index, without a local SVD recomputation, but using the metadata at each node for coordinate transformation and reducing the number of dimensions. The SVD computation is invoked periodically, when the recall for k-NN queries drops below a certain threshold.

The approximate algorithm for k-NN queries presented in [15] may result in unacceptably small values of recall for higher values of the NMSE. This issue is dealt with in [5] by conducting an offline experiment to determine a sufficiently large $k^* > k$ for issuing the k-NN query to yield an acceptable value of recall (see Section 3). In this paper we propose an exact algorithm to process k-NN queries on dimensionality reduced clusters produced by CSVD. The algorithm is an extension of the algorithm specified in [10,11], which is based on the lower-bounding property [10,5,11]. Experimental results with two datasets show that the new algorithm requires less CPU time than the approximate algorithm, which is issued with a known value of k^*.

The paper is organized as follows. Section 2 describes the CSVD method and the k-NN search algorithm. Experimental results are given in Section 3, followed by conclusions in Section 4.