Chapter 1

Introduction

A dimensional data point is often be used to describe a real world object. Color histograms can be used to present a image. For searching a document, dimensionality consisting of axes corresponding to keys is used to depict documents. A Shogi(or chess) board can be considered as a data point in high dimensional space.

If an object corresponds to a point of dimensional space, the dissimilarity between two objects is reflected as the distance of two corresponding data points. Minkowski L_p distance is frequently used. Many applications deal with large amounts of dimensional data.

1.1 Applications of High Dimensional Data

- *Similarity search:* Retrieval of multimedia objects (such as images) is usually by using features extracted from those objects. Color histogram, shape descriptors and texture vector are always be used as features. Similarity search is generally divided

into two types. They are k-NN search and r-neighbor search. k-NN search is used to pick up k-th nearest objects from a query point, while r-neighbor search is to find out the objects whose distance from the query point is less than r. In order to make the answer more practical, it is important to choose distance function so as to capture the human perception of "similarity". Similarity search appear in many application fields such as e-commerce(e.g. find all shoes in the shopping catalog similar to given shoes), medical diagnosis or research and computer aided design.

- *Time series database retrieval:* Similarity search in time series data is useful as a tool of end user analysis. Using transformation(e.g Discrete Fourier Transform(DFT)[3, 19], Discrete Wavelet Transform(DWT)[15, 28], Singular Value Decomposition(SVD) [15], time series segments are converted to dimensional data points. Applications include searching by a doctor for a particular pattern in the ECG database for prediction etc[27].

- *Visualization:* Each dimensional data record contains values for several attributes which together define a dimensional space. For example, in the student database, each student record contains information on age, weight and marks of each subject etc. Visualization application finds interesting clusters in visualizing all students as points in 2 or 3-dimensional space.

Work on dimensional index structures dates back to early 80s. The first multidimensional index structures proposed were the spatial index structures, for example, R-tree[23], KDB-tree[32], grid-tree[31]. Although the above index structures work well in the low

dimensional spaces(2-3 dimensions) which they are designed for. They are not efficiently applied to high dimensional spaces. A simple sequential scan through the entire dataset to answer the query is often faster than accessing the data using a spatial access method[38]. In order to be applied for high dimensional data, dimensionality reduction is a promising solution. In this dissertation, we propose several approaches of dimensionality reduction techniques as discussed in the next section.

1.2 Approaches

Dimensionality Reduction in L_1 Metric Space: We investigate methods for retrieving objects within some distance from a given object by utilizing Spatial Access Methods(SAMs) R-tree and its family, which usually assume Euclidean metric. In this thesis, we attempt to apply SAMs to L_1 metric. First, we prove that objects in discrete L_1 (or Manhattan distance) metric space can be embedded into vertices of a unit hyper-cube when the square root of L_1 distance is used as the distance. To take full advantage of R-tree spatial indexing, we have to project objects into a space of relatively lower dimension. We adopt FastMap[18] by Faloutsos and Lin to reduce the dimensionality of object space. The range corresponding to a query (Q, h) for retrieving objects within distance h from an object Q is naturally considered as a hyper-sphere even after FastMap projection, which is an orthogonal projection in Euclidean space. However, it is turned out that the query range is contracted into a smaller hyper-box than the hyper-sphere by applying FastMap to objects embedded in the above mentioned way. Finally, we give a brief summary of experiments

in applying our method to Japanese chess boards.

High-Dimensional Index Structure: High dimensional data often are not uniformly distributed[4, 5, 16]. In order to exploit such skewed property, we propose a new dimensionality reduction technique and an indexing mechanism for high dimensional datasets. The proposed technique reduces the dimensionality whose coordinates are less than a critical value with respect to each data point. One of the advantages of the proposed technique is that it reduces the dimensionality *datawise*. This flexible dimensionality reduction contributes to improve indexing mechanisms for high dimensional datasets whose coordinate histogram has the tendency of Zip distribution. To apply the proposed technique to information retrieval, CVA-file (Compact VA-File) which is a revised version of the VA-file is developed. The size of data points stored in index files is reduced. The effectiveness is confirmed by synthetic and real data.

Time Series Databases Techniques: Similarity search in large time series databases poses several new indexing challenges. It is a difficult problem because of the typically high dimensionality of the raw data, for example, the raw ECG data in [27] has dimensionality between 256 and 1024. Because of the high dimensional nature of the data, the difficulties associated with dimensionality curse arise. The most promising solution involves performing dimensionality reduction on data, then indexing the reduced data with a dimensional index structure. In this work we introduce a new approach called grid-based Datawise Dimensionality Reduction(DDR) which attempts to preserve the characteristics of time series. We then apply quantization to construct an index structure.

Parametric Visualization: Visualization is one of effective methods for analyzing the way in which the high dimensional data are widely scattered. Dimensionality reduction, such as PCA, can be used to map high dimensional data to 2 or 3-dimensional space. For large dimensional data, FastMap is practical because of its linear computation complexity. We propose an algorithm *HyperMap* and develop a novel parametric visualization. Our algorithm can be seen as a generalization of FastMap. It holds the linear computation complexity, and overcomes several main shortcomings. First, the target space is flexible, and how data are scattered can be observed in various viewpoints. Furthermore, the restriction that each dimension has only two pivot objects is released. Then in visualization, the number of pivot objects can go beyond the limitation 6. This reduces the impacts of selecting a bad pivot object.

The rest of this thesis is organized as follows. Chapter 2 introduces an L_1 metric index structure. Chapter 3 describes a datawise dimensionality reduction technique, and applies it to VA-file. Chapter 4 proposes an indexing structure for large time series. Chapter 5 describes an interactive data mining tool called parametric visualization. In Chapter 6, we summarize the contribution of this thesis and outline some directions for future research.