# Chapter 6

# Conclusion and Future Work

We conclude this dissertation with a summary of our contributions and directions for future work.

## 6.1 Summary

For managing large high dimensional datasets, we proposed several dimensionality reduction techniques and several index structures.

- **Dimensionality Reduction in $L_1$ Metric Space:** Using spatial access methods (SAMSs) is an efficient technique of similarity search. It is no trivial to define a distance function that best reflects human perception regarding similarity measurements. Unlike most state-of-the-art technique of indexing high dimensional data which the Euclidean distance is adopted. We attempted to index high dimensional data with $L_1$ distance function, because in some cases, the distance function $L_1$ is adapted well

to distinguishing those objects. For scaling index structure, the technique of dimensionality reduction was adapted in the $L_1$ metric space. We found an interesting theorem that $L_1$ distance range from a query point can be accurately reflected into the Euclidean space $L_1$ metric embedded.

- **High-Dimensional Index Structure:** The index mechanisms based on tree structure are decline when the dimension of data become beyond $5 \sim 10$, because the most of nodes in tree are needed to scanned, that is, the power of pruning lost. The sequential scanning is advantageous for high dimensional dataset due to its efficient sequential I/O accesses. VA-file method makes the best use of advantages of sequential scanning, a two steps retrieval technique was proposed. However, VA-file has the same drawback like sequential scanning, the number of I/O access is proportion to the number of dimensions. Form the observation that the coordinates of high dimensional data is skewed, we proposed a novel method of dimensionality reduction, which operates reduce dimensions data by data. The distance information lost by dimensionality reduction reaches minimum. We apply the datawise dimensionality reduction to VA-file, a new version CVA-file is developed.

- **Time Series Databases Techniques:** Similarity search in time series databases is a difficult problem due to the typically high dimensionality of the raw data. The most promising solution involves performing dimensionality reduction on the data, then indexing the reduced data with a dimensional index structure. We proposed a dimensionality reduction technique aggregating before and behind data to an approx-

imation. the reduced representation closely approximates the original signal. A new time series index structure called DDR is proposed. It has significantly decreased I/O access than existing index structure.

- **Parametric Visualization:** Clustering high dimensional reduction is a challenge work, because high dimensional space has high degree of freedom, data points could be so scattered that every distance between them might yield no significant difference. An interactive clustering tool is desirable. Furthermore, for large datasets, the computation complexity over linear is always not available. We developed a clustering tool that end-user can change the viewpoints by tuning parameter. The kernel algorithm is called HyperMap that is a generalization of FastMap. It preserves the linear computation complexity of FastMap.

## 6.2 Future Works

There are several interesting directions of future work based on the work described in this thesis.

- *sequence data:* In this thesis, we proposed datawise dimensionality reduction and indexing techniques. These can be applied to sequence data as well. Examples of sequence data includes gene/protein sequences and stream data. Developing new search and mining techniques for such types of data based on adaptive representations is an interesting direction of research.

- *visual data mining:* In proposed parametric visualization, how to determine the number of pivot objects for each hyperaxis is a future work.

- *using HyperMap for indexing high dimensionality dataset:* We developed a parametric visualization using mapping high dimensional data to lower display space. We are considering how to index the target space mapped by HyperMap.

- *non-linear dimensionality reduction:* Classical linear dimensionality reduction methods such as PCA have limitation in high dimensional datasets. We think that non-linear dimensional reduction methods are promise techniques for indexing and visualization for high dimensional datasets.