Index Method Based on Dimensional Reduction

Doctoral Program in Engineering University of Tsukuba

2003.3

Jiyuan An
Acknowledgements

First of all, I would like to devote my Ph.D dissertation to the most important person in my life: my mother Yundi Yuan. I can not imagine a person suffering more hardships than her. Twenty-one years ago she left the world due to overwork at the age of only 45 years. She was illiterate, but she had a great desire that her children should be capable of reading and writing. I am grateful to my wife Qinying and our child Bairu for giving me the support I needed to finish my Ph.D.

My advisor, Professor Nobuo Ohbo, gave me the freedom to explore new topics, and at the same time he provided numerous insightful suggestions. I am very fortunate to have met such an open-minded advisor: he accepted me as his student even though he knew that I had a very different background.

I am very thankful to the fact that I have had another two advisors, Assistant Professor Hanxiong Chen and Assistant Professor Kazutaka Furuse. This gave me the unusual advantage of double resources that made my Ph.D study most efficient.

I would like to thank Professor Hiroyuki Kitagawa who gave me many precious suggestions and comments. I wish to thank Assistant Professor Jeffrey Yu of the Chinese
University of Hong Kong who has always been encouraging and helpful to me.

I must thank my Master course advisor Takeshi Shinohara of the Kyushu Institute of Technology: He introduced me to the field of study and opened up a wide range of topics, and he let me know interesting of study.
ABSTRACT

INDEX METHOD BASED ON DIMENSIONAL REDUCTION

High-dimensional data, such as documents, digital images, and audio clips, can be considered as spatial objects. The distances in a feature space between two objects measure their dissimilarities, and, the spatial indexing/access method R-tree [23] and its family on the space can be applied to the problem of the approximate retrieval. However, how to define the distance function is an important problem in high dimensional datasets. Though Euclidean distance ($L_2$) is commonly used, in some cases the metric other than $L_2$ is more appropriate for describing the feature of the data. However, except for $L_2$ distance function, spatial index method R-tree and its family are not applicable, because they are based on Euclidean space. In this dissertation, we propose a way to map $L_1$ metric to a Euclidean space, then R-tree is applied to the Euclidean space.

In many applications of dimensional datasets, such as, content-based retrieval, similarity search and data mining for time series, the space in which objects embedded has usually high dimensionality (hundreds - thousands). Most dimensional index structures proposed so far do not practical beyond 10-15 dimensional spaces because of so-called 'dimensionality curse'. The effectiveness of R-tree is based on pruning most of branches at every level of a tree. Although the random access used in the R-tree scheme is less effective than the sequential access, its defect is compensated by discarding unnecessary data. However, when the number of dimensions becomes higher, the overlapping between branches of a
tree increases rapidly, and most of branches are needed to be accessed. As a result, a simple sequential scan through the entire dataset to answer the query is often faster than using a tree dimensional index structure. To break the curse of dimensionality, the 2-step retrieval method VA-file [38] was proposed. In this scheme, a compressed data is scanned linearly in the first phase and a small set of candidates are extracted. In the second phase the answer is picked out with relatively fewer accesses to original data file. VA-file is based on the effectiveness of the sequential access, and it has extremely good performance in uniformly distributed data. However, by observing the real high dimensional datasets, we found that their coordinates histograms have tendency of Zipf's distribution. In this dissertation, a Compact VA-file (CVA-file) is proposed to make VA-file adapted to various kinds of real dataset.

Because of the sparseness of high dimensional space, data mining for high dimensional datasets is a challenging work. Because it is difficult to analyses the order of scattering of data in a high dimensional space, visualization in which data are mapped into 2 or 3-dimensional space, is an efficient method. Most visualization methods proposed so far use a fixed target space where end-user can see the distribution of data from only one viewpoint. In these scheme, it frequently happens that since many distinct clusters are mapped into one large area, user cannot distinguish each cluster though the visual image. Moreover, Methods based on Principal Component Analysis (PCA) consume a large amount of computation time. However, for large dimensional datasets, the linear time complexity is disable. We develop an interactive visualization method by using a novel mapping method
called HyperMap. By tuning parameters, the order of scattering of data in target space can be changed. End-user can extract clusters in the step-by-step fashion. Furthermore, HyperMap algorithm has the linear time complexity. Its effectiveness is confirmed by synthetic and real dataset.
Contents

Acknowledgements iii

Abstract v

Contents viii

1 Introduction 1

1.1 Applications of High Dimensional Data 1

1.2 Approaches 3

2 Approximate Retrieval of High-dimensional data with $L_1$ Metric by Spatial Indexing 6

2.1 Introduction 6

2.2 Embedding $L_1$ distance into Euclidean space 11

2.3 Contraction of Query Range by FastMap 12

2.4 Experimental Results

   – Approximate Retrieval of Japanese Chess Boards 17
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.1</td>
<td>Distance between boards</td>
<td>17</td>
</tr>
<tr>
<td>2.4.2</td>
<td>FastMap projection and R-tree spatial indexing</td>
<td>18</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Effect of contraction of query range by FastMap</td>
<td>18</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Approximate retrieval of boards</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>Concluding Remarks</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td><strong>CVA-file: An Index Structure for High-Dimensional Datasets</strong></td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>The CVA-file technique</td>
<td>26</td>
</tr>
<tr>
<td>3.2.1</td>
<td>VA-file</td>
<td>26</td>
</tr>
<tr>
<td>3.2.2</td>
<td>VA-file in KLT domain</td>
<td>27</td>
</tr>
<tr>
<td>3.2.3</td>
<td>CVA-file</td>
<td>28</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Estimating Bounds in CVA-file</td>
<td>33</td>
</tr>
<tr>
<td>3.2.5</td>
<td>CVA-file Algorithm</td>
<td>36</td>
</tr>
<tr>
<td>3.3</td>
<td>Performance Evaluation</td>
<td>37</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Real Dataset</td>
<td>40</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Synthetic Dataset</td>
<td>44</td>
</tr>
<tr>
<td>3.4</td>
<td>Conclusions</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td><strong>Grid-Based Indexing for Large Time Series Databases</strong></td>
<td>48</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>48</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Related Work</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Approach</td>
<td>52</td>
</tr>
</tbody>
</table>
4.2.1 Notation and Terminology .................................. 52
4.2.2 Overview of our Approach .................................. 53
4.3 Data Representation ............................................. 55
  4.3.1 The DDR representation .................................. 55
  4.3.2 Distance Measures Defined for DDR ..................... 59
4.4 Indexing DDR ................................................. 62
4.5 Experimental Evaluation ....................................... 65
  4.5.1 Experimental Result: Reduction of Dimensionality ...... 66
  4.5.2 Experimental Result: Comparison on number of page accesses ... 68
4.6 Conclusions .................................................. 71

5 HyperMap: Parametric Linear Visualization for High-Dimensional Data Clustering 72
  5.1 Introduction ................................................ 72
  5.2 Visualizing Large High-Dimensional Data in a 3-Dimensional Space .... 76
  5.3 HyperMap ................................................ 77
    5.3.1 FastMap ............................................. 77
    5.3.2 HyperMap Overview .................................. 80
    5.3.3 Pivot Selection .................................... 81
    5.3.4 Computing Relative Coordinate ....................... 82
    5.3.5 Computing Projected Distance in the Complementary Space .... 86
    5.3.6 Hypercoordinate Calculation ......................... 87
5.3.7 Algorithm of HyperMap .................................. 87
5.4 Empirical Results ........................................... 89
  5.4.1 Synthetic Data Generation .............................. 89
  5.4.2 Getting Accurate Results by Tuning Weight W ...... 90
  5.4.3 Real dataset ........................................... 91
5.5 Conclusions .................................................. 92

6 Conclusion and Future Work ................................. 94
  6.1 Summary .................................................. 94
  6.2 Future Works ............................................. 96

A Appendix ....................................................... 98
  A.1 Hypercoordinate in 2-hyperaxis .......................... 98
  A.2 The number of pivot object from \(k\) to \(k + 1\) .......... 98

Bibliography ...................................................... 101