High-Dimensional Index Structures Database Support for Next Decade's Applications

Stefan Berchtold
AT&T Labs Research
Florham Park, NJ
berchtol@research.att.com

Daniel A. Keim
University of Halle-Wittenberg
Germany
keim@informatik.uni-halle.de

1 Introduction

During recent years, a variety of new database applications has been developed which substantially differ from conventional database applications. For example, new database applications such as data warehousing produce very large relations which require a multidimensional view on the data, and in areas such as multimedia and CAD a content-based search is essential which is often implemented using some kind of feature vectors. All the new applications have in common that the underlying database system has to support the processing of queries on large amounts of high-dimensional data. Now, we may ask what the difference is between processing low- and high-dimensional data. A result of recent research activities is that basically none of the querying and indexing techniques which provide good results on low-dimensional data also performs sufficiently well on higher-dimensional data. The problem of dealing with high-dimensional spaces has therefore been addressed in a variety of recent database research projects. The goal of the tutorial is to spread the knowledge about high-dimensional spaces and the proposed techniques to a large community of both, researchers and practitioners — researchers who are interested in querying and indexing techniques for high-dimensional data, and practitioners who are interested in the state-of-the-art of database support for their applications.

The tutorial is structured as follows: In the first section, we describe two examples of new database applications which demonstrate the need for efficient query processing techniques in high-dimensional spaces. In the second section, we discuss the effects occurring in high-dimensional spaces — first from a pure mathematical point of view and then from a database perspective. Next, we describe the different approaches for modeling the costs of processing queries on high-dimensional data. The description of the different approaches demonstrates nicely what happens if we ignore the special properties of high-dimensional spaces. In the fourth section, we then provide a structured overview of the proposed querying and indexing techniques, discussing their advantages and drawbacks. In this section, we also cover a number of additional techniques dealing with optimization and parallelization. In concluding the tutorial, we try to stir further research activities by presenting a number of interesting research problems.

2 Outline

1. **Modern Database Applications**
   - Multimedia and CAD Databases
   - Data Warehouses
2. **Effects in High-Dimensional Spaces**
   2.1. Mathematical Background
   2.2. Database Issues (Query Selectivity, Data Distributions)
   2.3. Indexing Issues (Split Strategies, Shape of Data Pages)
3. **Models for High-Dimensional Query Processing**
   3.1. Traditional Model
   3.2. Exact Model
   3.3. Analytical Model
4. **Indexing High-Dimensional Spaces**
   4.1. Criterions for Classification
   4.2. k-d-Tree-based Techniques
   (k-d-Tree, VAM-Split Trees, LSD^h-Tree)
   4.3. R-Tree-based Techniques
   (R-Tree, X-Tree, SS-Tree, SR-Tree)
   4.4. Other Techniques
   (TV-Tree, Voronoi-based Indexing, Pyramid-Technique)
   4.5. Optimization and Parallelization
   (Tree-Striping, Parallel Declustering, Optimizing Data Space Partitioning)
5. **Open Research Topics**
   5.1. Partitioning Strategies
   5.2. Parallel Query Processing
   5.3. Data Reduction
6. **Summary and Conclusion**

3 Tutorial Notes

http://www.informatik.uni-halle.de/~keim/SIGMOD98Tutorial.ps