Tutorial 5: Indexing High-Dimensional Spaces: 
Database Support for Next Decade’s Applications

Stefan Berchtold and Daniel A. Keim  
sth gmbh, University of Halle, Germany

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. Also, the tutorial will be very interesting for non-database computer scientists because the problem of dealing with high-dimensional spaces has a large number of other applications such as robot motion planning, optimization problems, and visualization techniques. Therefore, a large part of the tutorial is dedicated to convey the understanding of the effects occurring in these spaces.

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.

Bio

Daniel A. Keim is working in the areas of indexing high-dimensional data spaces, data mining, and data visualization. He developed several new techniques for improving the access to high-dimensional data and applied them in application areas such as similarity search and data mining. Dr. Keim has published extensively on indexing high-dimensional data spaces, data mining, and data visualization. Additional research interests include nearest-neighbor search, similarity search in CAD and medical databases, clustering, and information visualization.

Dr. Keim received his diploma (equivalent to an MS degree) in Computer Science from the University of Dortmund in 1990, his Ph.D. in Computer Science from the University of Munich in 1994, and his habilitation degree from the University of Munich in 1997. Currently, he is associate professor at the Institute for Computer Science of the University of Halle, Germany.
Stefan Berchtold is working in the areas of data mining, indexing high-dimensional data spaces, similarity search and data visualization. He authored several papers and patents in these areas. Dr. Berchtold received a diploma in Computer Science from the Technical University of Munich in 1993. In 1994, he worked as a consultant in industry. Then, he joined the University of Munich and received a Ph.D. in Computer Science in 1997. From 1997 to 1998, he was with the AT&T Laboratories in Florham Park working as a Senior Researcher. Currently, he is CEO of stb gmbh, a German start-up company that is developing database applications and does consulting in the area of data mining.