Chapter 4
Compression Schemes of High Dimensional Data for MOLAP

K. M. Azharul Hasan
Khulna University of Engineering and Technology (KUET), Bangladesh

ABSTRACT

The exploration of the possibility of compressing data warehouses is inevitable because of their non-trivial storage and access costs. A typical large data warehouse needs hundreds of gigabytes to a terabyte of storage. Performance of computing aggregate queries is a bottleneck for many Online Analytical Processing (OLAP) applications. Hence, data warehousing implementations strongly depend on data compression techniques to make possible the management and storage of such large databases. The efficiency of data compression methods has a significant impact on the overall performance of these implementations. The purpose of this chapter is to discuss the importance of data compression to Multidimensional Online Analytical Processing (MOLAP), to survey data compression techniques relevant to MOLAP, and to discuss important quality issues of MOLAP compression and of existing techniques. Finally, we also discuss future research trends on this subject.

INTRODUCTION

Data compression is widely used in data management to save storage space and network bandwidth. The main benefits that are achieved in data compression are well described by Bassiouni 1985 for different contexts and applications. The most obvious advantage of data compression is that of reducing the storage requirement for the database. Reducing the storage requirement of databases is equivalent to increasing the capacity of the storage medium. Since compressed data are encoded using a smaller number of bytes, transfer of compressed information from one place to another requires less time and hence results in a higher effective transfer rate. Since data compression reduces the loading of I/O channels, it becomes feasible to process more I/O requests per second and hence achieve higher effective channel utilization. Most importantly, however, is the application of data compression

DOI: 10.4018/978-1-60566-816-1.ch004

in reducing the cost of data communication in distributed networks. In order to use or interpret compressed data, it is necessary to restore the information to its uncompressed format. To do this, a decoding algorithm must be available, and performance concerns are relevant for that operation. In some applications, data compression can also lead to other types of improvement in system performance. For example, in some index structures it is possible through compression to pack more keys into each index block. When the database is searched for a given key value, the key is first compressed and the search is performed against the compressed keys in the index blocks. The net effect is that fewer blocks have to be retrieved and thus the average search cost is reduced.

On-line Analytical Processing (OLAP) is a database acceleration technique used for deductive analysis. The main objective of OLAP is to have constant-time or near constant-time answers for many typical queries. There are two types of OLAP, namely ROLAP (Relational OLAP) and Multidimensional Online Analytical Processing (MOLAP). In ROLAP, the data is usually stored in the form of "summary tables". ROLAPs are built on top of standard relational database systems, whereas MOLAPs are based on multidimensional database systems. The data structures in which ROLAPs and MOLAPs store datasets are fundamentally different. ROLAPs use relational tables as their basic data structure and MOLAPs store their datasets as multidimensional arrays. Those large multi-dimensional arrays are used as basic data structures for scientific computations, business analysis, and visualization, where huge amounts of data manipulation are necessary. The multi-dimensional rectangular arrays, both dense and sparse depending on the context, form the fundamental abstract data structure used in different computation schemes. One area where multidimensional arrays are commonly used is data warehousing and Online Analytical Processing (OLAP), which often requires extraction of statistical information for decision support.

In MOLAP applications, data compression is important because database performance strongly depends on the amount of available memory. A MOLAP is a set of *multidimensional datasets* and is designed to allow for the efficient and convenient storage and retrieval of large volumes of data that is closely related, viewed and analyzed from different perspectives. The multidimensional arrays that are linearized to store multidimensional datasets normally have high degree of sparsity and need to be compressed. It is therefore desirable to develop techniques that can access the data in their compressed form and can perform logical operations directly on the compressed data. Multidimensional arrays are good to store dense data, but most datasets are sparse, which wastes huge memory, since a large number of array cells are empty and thus are very hard to use in actual implementations. In particular, the sparsity problem increases when the number of dimensions increases. This is because the number of all possible combinations of dimension values increases exponentially, whereas the number of actual data values would not increase at such a rate. Efficient storage schemes are required to store such sparse data for multidimensional arrays for MOLAP implementations. In this chapter, a survey of the compression schemes for multidimensional data is presented. The data compression techniques are not only important for data warehousing implementation but also for any kind of large database implementation such as Scientific and Statistical Databases (SSDB).

Some of the most relevant issues concerning data compression are: the ability to perform efficient and random searching in compressed databases for a given logical position in the original database; and then the ability to provide an efficient mapping from arbitrary positions in the compressed data back to the corresponding logical position in the original database.