OLAP Data Cube Compression Techniques: A Ten-Year-Long History

Alfredo Cuzzocrea

ICAR-CNR and University of Calabria
I-87036 Rende, Cosenza, Italy
cuzzocrea@si.deis.unical.it

Abstract. *OnLine Analytical Processing* (OLAP) is relevant for a plethora of *Intelligent Data Analysis and Mining Applications and Systems*, as it offers powerful tools for exploring, querying and mining massive amounts of data on the basis of fortunate and well-consolidated multidimensional and a multi-resolution metaphors over data. Applicative settings for which OLAP plays a critical role are manyfold, and span from *Business Intelligence* to *Complex Information Retrieval* and *Sensor and Stream Data Analysis*. Recently, the Database and Data Warehousing research community has experienced an explosion of OLAP-related methodologies and techniques aimed at improving the capabilities and the opportunities of complex mining processes over heterogeneous-in-nature, inter-related and massive data repositories. Despite this, open problems still arise, among which the so-called *curse of dimensionality problem* plays a major role. This problem refers to well-understood limitations of state-of-the-art OLAP data processing techniques in elaborating, querying and mining multidimensional data when data cubes grow in size and dimension number. This evidence has originated a large spectrum of research efforts in the context of *Approximate OLAP Query Answering* techniques, whose main idea consists in *compressing target data cubes in order to originate compressed data structures able of retrieving approximate answers to OLAP queries at a tolerable query error*. This research proposes an excerpt of a ten-year-long history of OLAP data cube compression techniques, by particularly focusing on three major results, namely $\Delta - Syn$, K_{LSA} and $\mathcal{LCS} - Hist$.

1 Introduction

OnLine Analytical Processing (OLAP) [6] is relevant for a plethora of *Intelligent Data Analysis and Mining Applications and Systems*, as it offers powerful tools for exploring, querying and mining massive amounts of data on the basis of fortunate and well-consolidated multidimensional and a multi-resolution metaphors over data. Applicative settings for which OLAP plays a critical role are manyfold, and span from *Business Intelligence* to *Complex Information Retrieval* and *Sensor and Stream Data Analysis*. Recently, the Database and Data Warehousing research community has experienced an explosion of OLAP-related

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 751–754, 2010.

methodologies and techniques aimed at improving the capabilities and the opportunities of complex mining processes over heterogeneous-in-nature, inter-related and massive data repositories.

A significant issue in dealing with OLAP data processing and querying is represented by the so-called *curse of dimensionality problem* [1], which, briefly, consists in the fact that when size and number of dimensions of the target data cube increase, multidimensional data cannot be accessed and queried efficiently. Starting from this practical evidence, several *Approximate OLAP Query Answering* techniques have been proposed during the last years, with alternate fortune. The main idea of these techniques consists in computing *compressed representations* of input data cubes in order to evaluate time-consuming OLAP queries against them, thus obtaining *approximate answers*. Despite compression introduces some approximation in the retrieved answers, it has been demonstrated [2] that fast and approximate answers are perfectly suitable to OLAP analysis goals, whereas exact and time-consuming answers introduce excessive computational overheads that, in general, are very often incompatible with the requirements posed by an online computation for decision making, as a very large number of tuples must be accessed in order to retrieve the desired exact answers.

This research proposes an excerpt of a ten-year-long history of OLAP data cube compression techniques, by particularly focusing on three major results: (*i*) $\Delta - Syn$ [4], an *analytical synopsis data structure* that introduces a polynomial approximation technique for OLAP data cubes; (*ii*) K_{LSA} [3], which further extends the $\Delta - Syn$ proposal in order to provide *accuracy control* over compressed OLAP data cubes; (*iii*) $\mathcal{LCS} - Hist$ [5], a *histogram-based complex methodology* for compressing massive-in-size high-dimensional OLAP data cubes. In the following, we provide an overview on these OLAP data cube compression techniques.

$\Delta - Syn$: Analytical Synopsis Data Structures Supporting Polynomial Approximation of OLAP Data Cubes [4] $\Delta - Syn$ is a synopsis data structure for OLAP data cubes that is based on an innovative *analytical interpretation* of multidimensional data and the well-known *Least Squares Approximation* (LSA), which provides support for approximate aggregate query answering in OLAP. According to the $\Delta - Syn$ methodology, the input data cube is interpreted as *a set of data rows*, to which appropriate *Discrete Impulsive Distributions* are associated. The final synopsis data structure is obtained by approximating the *Cumulative Distribution Functions* of such distributions with a set of *polynomial coefficients* provided by the LSA method, and storing these coefficients instead of the original data. This allows us to achieve a compact representation of the original data cube, being the size of the polynomial coefficient set is bounded by the storage space $\mathcal{B}$ available for housing the compressed data structure. In this research, an efficient algorithm that takes the input data cube $\mathcal{D}$ and the storage space $\mathcal{B}$, and builds $\Delta - Syn$ with low spatio-temporal complexity is proposed. OLAP queries are issued on the compressed representation using an optimized ad-hoc procedure, thus reducing the number of disk accesses needed

to retrieve (approximate) answers. As demonstrated in [4], $\Delta - Syn$ provides good performance on both synthetic and real data cubes, even in comparison with other well-known compression techniques presented in literature, such as *histograms*, *wavelets* and *random sampling*.

K_{LSA}: Accuracy Control Over Compressed Data Cubes for Quality-of-Answer OLAP Tools [3] K_{LSA} is a state-of-the-art OLAP data cube compression technique that further extends the $\Delta - Syn$ proposal in order to provide *accuracy control* over compressed OLAP data cubes. K_{LSA} allows us to drive the compression process of data cubes in dependence on the accuracy required by external OLAP users/applications via determining the *degree of approximation* of final answers by means of meaningfully exploiting theoretical foundations offered by the LSA method. The main idea of the K_{LSA} proposal relies on two major assertions: (*i*) rigorously modeling and handling the degree of approximation of retrieved answers to OLAP queries over synopsis data structures; (*ii*) efficiently providing approximate answers having a desired accuracy, through setting the latter as an input parameter of the entire LSA-based compression process. This results in a novel *parametric LSA method* that meaningfully extends the baseline method and allows us to introduce the so-called *accuracy-aware LSA compression technique*. Given an input data distribution, the parametric LSA method is able of computing the *best* approximating function for this distribution in dependence on a *fixed* (i.e., required) accuracy. This baseline task is in turn exploited by the accuracy-aware LSA compression technique to achieve *accuracy-aware* compressed OLAP data cubes. A secondary-but-relevant contribution of the K_{LSA} proposal is represented by some effective optimizations of the above-sketched data cube compression technique that allow higher effectiveness and higher compression ratios to be achieved. Most importantly, K_{LSA} enables the design and the development of next-generation *Data Warehousing and OLAP Server Systems*, called *Quality-of-Answer (QoA) OLAP Tools*, which introduce an innovative paradigm according to which OLAP users/applications and DW servers implement an *application protocol* such that the final degree of approximation of retrieved answers is established by trading-off the required accuracy and the amount of storage space available for housing the compressed representation of the target data cube. A comprehensive experimental campaign on the K_{LSA} performance in compressing OLAP data cubes and supporting approximate query answering on so-compressed data cubes against several kinds of synthetic multidimensional data sets clearly demonstrates the superiority of K_{LSA} over significant similar techniques [3].

$\mathcal{LCS} - Hist$: Scalable Histogram-based Approximation of Massive-In-Size High-Dimensional OLAP Data Cubes [5] The $\mathcal{LCS} - Hist$ proposal introduces a *histogram-based complex methodology* for compressing massive-in-size high-dimensional OLAP data cubes whose main goal consists in overcoming actual *scalability limitations* of popular histogram-based data cube compression approaches. Indeed, classical histograms perform well on small-in-size low-dimensional data cubes whereas they do not scale satisfactorily on massive

high-dimensional data cubes. For this reason, when the latter kind of data cubes are considered, we generally observe a significant performance degradation in both representing the input data domain and introducing low (query) errors in the retrieved approximate answers. To adequately face-off this drawback, the methodology underlying $\mathcal{LCS}-Hist$ defines an innovative data cube compression methodology that combines a *collection* of intelligent multidimensional data modeling and processing techniques: (*i*) $\mathcal{L}$*inear programming*, (*ii*) $\mathcal{C}$*onstrained partitions of multidimensional data domains*, and (*iii*) $\mathcal{S}$*imilarity metrics on one-dimensional histograms*. The main motivation of the novel vision carried out by $\mathcal{LCS}-Hist$ is the following. Since traditional histogram-based data cube compression techniques expose problematic limitations when applied to massive high-dimensional data cubes, combine intelligent multidimensional data modeling and processing techniques in order to obtain a final compressed data structure, the multidimensional histogram $\mathcal{LCS}-Hist$, that, although paying something in terms of computational overheads, allows us to achieve excellent performance in both representing the input data cube and efficiently supporting approximate query answering to resource-intensive OLAP queries against the compressed data structure. As proven in the experimental evaluation and analysis provided in [5], contrary to state-of-the-art histogram-based data cube compression techniques, $\mathcal{LCS}-Hist$ ensures high scalability and efficiency on massive high-dimensional data cubes, the most common kind of data cubes one can find in real-life OLAP applications.

References

1. Berchtold, S., Bhm, C., Kriegel, H.-P.: The pyramid-technique: Towards breaking the curse of dimensionality. In: Proceedings of the 1998 International Conference on Management of Data (SIGMOD 1998), pp. 142–153 (1998)
2. Cuzzocrea, A.: Overcoming limitations of approximate query answering in olap. In: Proceedings of the 9th International Symposium on Database Engineering and Applications (IDEAS 2005), pp. 200–209 (2005)
3. Cuzzocrea, A.: Accuracy control in compressed multidimensional data cubes for quality of answer-based olap tools. In: Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM 2006), pp. 301–310 (2006)
4. Cuzzocrea, A.: Improving range-sum query evaluation on data cubes via polynomial approximation. Data & Knowledge Engineering 56(2), 85–121 (2006)
5. Cuzzocrea, A., Serafino, P.: LCS-hist: Taming massive high-dimensional data cube compression. In: Proceedings of the 12nd International Conference on Extending Database Technology (EDBT 2009), pp. 768–779 (2009)
6. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, A., Pellow, F., Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery 1(1), 29–53 (1997)