# Report on the Tenth ACM International Workshop on Data Warehousing and OLAP (DOLAP'07)

Torben Bach Pedersen
Department of Computer Science
Aalborg University
Selma Lagerløfsvej 300
DK-9220 Aalborg Ø, Denmark
tbp@cs.aau.dk

Il-Yeol Song
College of Information Science and Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104, USA
songiy@drexel.edu

## General Terms

Data Warehousing, On-Line Analytical Processing (OLAP)

## 1. INTRODUCTION

This paper presents an overview of DOLAP'07, the 10th ACM International Workshop on Data Warehousing and OLAP, held on November 9, 2007 in Lisbon, Portugal in conjunction with CIKM'07, the ACM 16th Conference on Information and Knowledge Management.

The mission of DOLAP is to explore novel research directions and emerging application domains in the areas of data warehousing and OLAP. Although, research in data warehousing and OLAP has produced important technologies for the design, management and use of information systems for decision support, there are still problems and research opportunities in the areas. Much of the interest and success in those areas can be attributed to the need for software and tools to improve data management and analysis given the large amounts of information that are being accumulated in corporate as well as scientific databases.

Nevertheless, the high maturity of these technologies as well as new data needs or applications not only demand more capacity or storing necessities, but also new methods, models, techniques or architectures to satisfy these new needs. Some of the hot topics in data warehouse research include distributed data warehouses, web warehouses, data streams, realtime DWs, GIS/location-based services, test and XML data, and biomedical data. Moreover, there are other aspects developed in other software areas such as security/privacy or quality, which still remain unexplored by current design methods or technologies for data warehouses

The call for papers attracted 28 submissions from Asia, Canada, Europe, and the United States. The program committee accepted 12 papers, yielding an acceptance rate of 42.9%. The papers were organized into four different sessions: 1) data warehouse design, 2) physical data organization, 3) data warehouse processing, and 4) spatio-temporal data warehouses and data mining. Finally, since DOLAP is the premier venue for data warehouse and OLAP research, the program included a research challenges panel session "Research Challenges for DW and OLAP Seen from Industry and Academia," where research challenges were proposed by four researchers and practitioners.


*SIGMOD Record*


## 2. DATA WAREHOUSE DESIGN

Romero et al. [1] presented a new approach to automate the multidimensional design of Data Warehouses. The approach proposed a semi-automated method that tried to find business-related multidimensional concepts from a domain ontology representing different (and potentially heterogeneous) data sources of a given business domain. The method identified common "business" multidimensional concepts from heterogeneous data sources for which the only common assumption was that they were described by an OWL ontology.

Song et al. [2] presented the SAMSTAR method which semi-automatically generates star schemas from an Entity Relationship Diagram (ERD), by analyzing both the semantics and the structure of the ERD. This eases the popular approach of developing star schemas based on existing ERDs using some heuristics. The novel features of SAMSTAR were (a) the use of the notion of Connection Topology Value (CTV) for identifying fact and dimension candidates and (b) the use of Annotated Dimensional Design Patterns (A_DDPs) as well as WordNet to extend the list of dimensions. The method was illustrated by applying it to examples from existing literature, showing that the outputs of SAMSTAR method are a superset of those of the existing methods.

Aouiche et al. [3] compared five probabilistic techniques for aggregate view size estimation. They observed that many available techniques for view-size estimation make particular statistical assumptions and that their errors can be large. In comparison, "unassuming" probabilistic techniques are slower, but the estimates are more accurate and reliable for very large view sizes, and these techniques use little memory. The paper compared five unassuming hashing-based view-size estimation techniques including Stochastic Probabilistic Counting and LOGLOG Probabilistic Counting. The experiments showed that only Generalized Counting, Gibbons-Tirthapura, and Adaptive Counting provide universally tight estimates irrespective of the size of the view; of those, only Adaptive Counting remains constantly fast as the memory budget is increased.

## 3. PHYSICAL DATA ORGANIZATION

Otoo et al. [4] presented techniques for optimal chunking of large multidimensional arrays which are commonly used in scientific computations as well as MOLAP. They investigated the problem of what shapes of array chunks give the minimum expected number of chunks over a query workload. The paper improved on a previous paper by Sarawagi and Stonebraker, by developing exact mathematical models of the problem and provide exact solutions using steepest descent and geometric programming methods. Experimental results with synthetic and real life workloads showed that the expected number of chunks are consistently within 2% of the true number of chunks.

Missaoui et al. [5] presented a probabilistic model for data cube compression and query approximation. The paper addressed the problem of automatically analyzing large multidimensional tables to get a concise and compact representation of data, identify patterns and provide approximate answers to queries. The paper analyzed the potential of a probabilistic modeling technique, called non-negative multiway array factorization, for approximating aggregate and multidimensional values. Using this technique, the set of components (clusters) that best fit the initial data set and whose superposition approximates the original data, was computed. The generated components was then be exploited for approximately answering OLAP queries such as roll-up, slice and dice operations. The proposed technique compared favorably to the log-linear modeling cube compression technique know from the literature for compression.

Stabno et al. [6] presented a technique of compressing bitmap indexes in data warehouses. The compression technique, called Run-Length Huffman (RLH), was based on both run-length encoding and Huffman encoding. RLH was implemented and experimentally compared to the Word Aligned Hybrid (WAH) bitmap compression technique that in the literature has been reported to provide the shortest query execution times. The experiments showed that RLH offers shorter query response times than WAH for certain cardinalities of indexed attributes. Moreover, bitmaps compressed with RLH are smaller than bitmaps compressed with WAH. Additionally, The authors proposed a modified RLH, called RLH-1024, which is designed to better support bitmap updates.

## 4. DATA WAREHOUSE PROCESSING

Tziovara et al. [7] presented techniques for deciding the physical implementation of ETL workflows. They dealt with the problem of determining the best possible physical implementation of an ETL workflow. As input, they provide a logical-level description of the ETL flow, and an appropriate cost model. They formulated the problem as a state-space problem and provided a suitable solution. They further extended this technique by intentionally introducing "sorter" activities in the workflow. This made it possible to search for alternative physical implementations with lower cost. They provided an experimental assessment of their proposal, based on a principled organization of test suites. The experiments showed that the intentional introduction of sorters can make the difference in the determination of the final solution in several cases.

Thiele et al. [8] presented techniques for partition-based workload scheduling in "living" (near-realtime) DW environments. Here, users expect both short response times for their queries and high data freshness. This is challenging due to the high loads and the continuous flow of write-only updates and read-only queries, which may be in conflict with each other. The paper thus presented the concept of Workload Balancing by Election (WINE), which allows users to express their individual demands on the Quality of Service and the Quality of Data, respectively. WINE applied this information in order to balance and prioritize over both queries and update transactions according to user needs. A simulation study showed that the proposed algorithm outperforms competitor baseline algorithms over the entire spectrum of workloads and user requirements.

Dehne et al. [9] considered the problem of efficient computation of view subsets. They argued that given the enormous size of the fact table in a star schema, virtually all current systems augment the primary fact table with a small number of focused summary tables (here called view subsets). Previous research already addressed the issue of the selecting of the most cost-effective summaries. However, the subsequent problem of actually efficiently computing a given view subset has received far less attention. The paper presented a suite of greedy algorithms for the construction of such view subsets. Experimental results demonstrated cost savings of between 20 and 70% relative to the naive alternative algorithms, depending upon the degree of materialization required.

## 5. SPATIO-TEMPORAL DATA WAREHOUSES AND DATA MINING

Escribano et al. [10] presented Piet, an implementation of a GIS-OLAP system. Piet made use of a novel query processing technique. First, a process called "sub-polygonization" decomposed each thematic layer in a GIS into open convex polygons. Second, another process then computed the so-called overlay of those layers, and stored in a database for later use by a query processor. The paper described the implementation of Piet. It also provided experimental evidence that overlay precomputation can outperform GIS systems that employ indexing schemes based on R-trees.

Kondratas et al. [11] presented CT-OLAP, a temporal multidimensional model and algebra for moving objects, focusing on so-called "sequenced queries", queries that are (conceptually) evaluated at each time instant, thus returning functions of time as a result. Applications like traffic analysis need support for sequenced queries on data about continuous changes, but current temporal OLAP technology does not support such queries since they are based on discrete time. The authors proposed a conceptual multidimensional model, CT-OLAP, that captures continuous functions of time. Its associated algebra supports sequenced analytical queries. CT-OLAP extends an existing powerful multidimensional model and algebra. CT-OLAP is currently being implemented in the Secondo DBMS. The work was motivated by requirements to a tool for traffic jam analysis formulated by the Municipality of Bozen-Bolzano, Italy.

Plantevit et al. [12] presented techniques for mining "unexpected" multidimensional rules. They argued that discovering unexpected rules is essential, particularly for industrial marketing applications. Much related work has been done for association rules, but none of it addresses sequences. The paper thus proposed techniques for discovering unexpected

multidimensional sequential rules in data cubes. It defined the concept of multidimensional sequential rule, and the notion of "unexpectedness." It formalized these concepts and defined an algorithm for mining such rules. Experiments on a real data cube showed the interestingness of the approach.

## 6. PANEL: RESEARCH CHALLENGES FOR DW AND OLAP SEEN FROM INDUSTRY AND ACADEMIA

Middelfart [13] presented thoughts on improving business intelligence speed and quality through the Observation-Orientation-Decision- Action (OODA) concept known from fighter pilot training. OODA was presented as a mean to identify three new desired technologies in business intelligence applications that could improve the speed and quality in the decision making processes. The desired techniques were 1) technologies that reduce the number of user interactions needed to cycle through an OODA loop, 2) technologies that can help users identify "sentinels," which are ideally measures that can give early warnings about a later influence on a business critical measure, and 3) Business Process Intelligence on the entire system of OODA loops.

Rizzi [14] proposed OLAP preferences as a research agenda. He argued that expressing preferences when querying databases is a natural way to avoid empty results and information flooding. It is also useful in general to rank results so that users may first see the data that better match their tastes. The paper outlined the main research issues to be faced in order to develop a system for handling user preferences on OLAP cubes.

Pedersen [15] presented challenges assocoated with "warehousing the world." He argued that DWs have become very successful in many enterprises, but only for relatively simple and "traditional" types of data. It is now time to extend the benefits of DWs to a much wider range of data, making it feasible to literally "warehouse the world". To do this, five unique challenges must be addressed: warehousing data about the physical world, integrating structured, semi-structured, and unstructured data in DWs, integrating the past, the present, and the future, warehousing imperfect data, and ensuring privacy in DWs

Sørensen [16] stated that "Even Straight Forward Data Warehouses are Complicated." From an industry point of view, he asked for research into problems like better tool support for dimensions, especially time dimensions, better tool integration across the whole set of BI tools, supporting fast, near-realtime updates of cubes and dimensions, and a standard, vendor-neutral, query language for cubes.

## 7. CONCLUSION

ACM DOLAP'07 was a highly successful event, with high-quality papers and very lively discussion. Now in its 10th year, DOLAP is alive and well, increasingly focusing on the wide range of novel challenges posed by the complex and dynamic nature of new types of data. The high quality of the papers is witnessed by the fact that the best papers of DOLAP'07 have been invited for a special issue of *Information Systems* for which they are currently under review.

## 8. ACKNOWLEDGMENTS

On behalf of the Program Committee we would like to thank all the authors of submitted papers for their interest in the workshop and the high quality of the submitted papers. We would also like to thank all the referees (both PC members and external reviewers) for their careful and dedicated work, both during the reviewing and the discussion phases. Working in cooperation with this program committee has been both a particular honor and a pleasure. Finally, we would like to express our gratitude to the members of the Organizing Committee of CIKM'07, the DOLAP Steering Committee, and our sponsors for their support in organizing this workshop.

## 9. REFERENCES

[1] O. Romero and A. Abello. Automating multidimensional design from ontologies. In [17], pp. 1–8, 2007.
[2] I.-Y. Song, R. Khare, and B. Dai. SAMSTAR: a semi-automated lexical method for generating star schemas from an entity-relationship diagram. In [17], pp. 9–16, 2007.
[3] K. Aouiche and D. Lemire. A comparison of five probabilistic view-size estimation techniques in OLAP. In [17], pp. 17–24, 2007.
[4] E. J. Otoo, D. Rotem, and S. Seshadri. Optimal chunking of large multidimensional arrays for data warehousing. In [17], pp. 25–32, 2007.
[5] R. Missaoui, C. Goutte, A. K. Choupo, and A. Boujenoui. A probabilistic model for data cube compression and query approximation. In [17], pp. 33–40, 2007.
[6] M. Stabno and R. Wrembel. RLH: bitmap compression technique based on run-length and huffman encoding. In [17], pp. 41–48, 2007.
[7] V. Tziovara, P. Vassiliadis, and A. Simitsis. Deciding the physical implementation of ETL workflows. In [17], pp. 49–56, 2007.
[8] M. Thiele, U. Fischer, and W. Lehner. Partition-based workload scheduling in living data warehouse environments. In [17], pp. 57–64, 2007.
[9] F. K. H. A. Dehne, T. Eavis, and A. Rau-Chaplin. Efficient computation of view subsets. In [17], pp. 65–72, 2007.
[10] A. Escribano, L. Gomez, B. Kuijpers, and A. A. Vaisman. Piet: a GIS-OLAP implementation. In [17], pp. 73–80, 2007.
[11] E. Kondratas and I. Timko. CT-OLAP: temporal multidimensional data model and algebra for moving objects. In [17], pp. 81–88, 2007.
[12] M. Plantevit, S. Goutier, F. Guisnel, A. Laurent, and M. Teisseire. Mining unexpected multidimensional rules. In [17], pp. 89–96, 2007.
[13] M. Middelfart. Improving business intelligence speed and quality through the OODA concept. In [17], pp. 97–98, 2007.
[14] S. Rizzi. OLAP preferences: a research agenda. In [17], pp. 99–100, 2007.
[15] T. B. Pedersen. Warehousing the world: a few remaining challenges. In [17], pp. 101–102, 2007.
[16] J. O. Sørensen. Even straight forward data warehouses are complicated. In [17], pp. 103–104, 2007.
[17] I.-Y. Song and T. B. Pedersen (Eds.). *DOLAP 2007, ACM 10th International Workshop on Data Warehousing and OLAP*, ISBN 978-1-59593-827-5, ACM Press, 2007.
[18] T. B. Pedersen (Ed.). *Information Systems* - Special Issue: Best paper of DOLAP'07, ISSN: 0306-4379, In preparation.