Report on the 9th International Workshop on Data Warehousing and OLAP (DOLAP 2006)

Panos Vassiliadis
Department of Computer Science,
University of Ioannina,
Ioannina, Hellas
pvassil@cs.uoi.gr

Il Yeol Song
College of Information Science and Technology
Drexel University,
Philadelphia, USA
song@drexel.edu

1 Introduction

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 area. Much of the interest and success in this area 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.

The *9th ACM International Workshop on Data Warehousing and OLAP* (*DOLAP'06*) was held in Arlington, VA, USA, in conjunction with the 15th ACM International Conference on Information and Knowledge Management (CIKM 2006), on November 10, 2006. Continuing the tradition of previous successful DOLAP workshops, the main aim of DOLAP 2006 workshop was to provide an international forum where both researchers and practitioners in the field of Data Warehousing and OLAP would share their findings in theoretical foundations, current methodologies, and practical experiences.

The call for papers attracted 26 submissions from Asia, Canada, Europe, and the United States. The program committee accepted 12 papers that were thematically grouped by the three major phases of a data warehouse lifecycle: design, construction and exploitation of a data warehouse. The conference was attended by about 16 people and involved lively discussions on the presented papers. More information on DOLAP 2006 (including presentations) is available at the web sites of the workshop: **http://www.cs.uoi.gr/~dolap06/** and **http://www.cis.drexel.edu/faculty/song/dolap.htm**.

2 Paper Presentations

Starting from the exploitation of the data warehouse, the accepted papers reflect the most recent trends in the research community and present novel *data mining and aggregation techniques* – quite a bit far from the traditional OLAP style of warehouse querying. All three papers on *data warehouse construction* involve real world cases and build upon them to provide novel design methods and vision for data warehouse design and construction. The papers in the area of *data warehouse design*, at the same time, explore hard topics such as ETL, indexing, view selection and spatial warehousing, which persist in attracting the interest of the research community. Finally, since DOLAP is the premier venue for data warehouse research, the program had to include papers with far-sighted, visionary ideas.

2.1 Advances in Data Warehousing

The presentations of the workshop started with a brush-stroke of vision: Prof. Timos Sellis presented an invited talk on the formal specification and optimization of ETL scenarios, whereas the program also hosted a paper on research directions for the field of data warehouse design.

In his talk entitled "*Formal specification and optimization of ETL scenarios*", *Timos Sellis* explained in detail the research area of Extract-Transform-Load (ETL) tools. ETL tools are responsible from populating the data warehouse with fresh data and clean data. To this end, a special-purpose ETL workflow deals with the extraction, transformation, cleansing and loading of the data on a regular basis. Timos presented novel results in the conceptual and logical modeling of such activities. The largest part of the talk, though, emphasized the optimization and tuning of the overall process, which involves the detection of opportunities for rearranging the ETL workflow for reasons of efficiency. The talk ended with ideas for future research that include ETL for non-traditional data, active data warehousing and richer semantics (possibly in terms of ontologies) to facilitate the optimization of the overall ETL process.

The paper by *Stefano Rizzi*, *Alberto Abelló*, *Jens Lechtenbörger*, *Juan Trujillo*, entitled "*Research in Data Warehouse Modeling and Design: Dead or Alive?*", was presented by Stefano Rizzi. In his talk, Stefano explained that the motivation for the paper was to report on the deliberations of the Dagstuhl seminar "Data Warehousing at the Crossroads" that took place in 2004 and specifically, on the parts that concern the issues related to data warehouse design. Concerning the conceptual modeling part, the main problem reported is the lack of a common standard for multidimensional modeling. Additional topics for research include design-for-security and design-for-data-mining. Logical modeling, at the same time, suffers from a semantic gap with conceptual modeling, resulting in research topics concerning design methods, quality metrics and interoperability issues concerning the warehouse metadata. Finally, a set of research issues concerning spatial, web, scientific and distributed warehouses was sketched.

2.2 Data mining & Approximation

This year, DOLAP 2006 differentiated from previous workshops as it includes papers related to knowledge extraction and approximation. The first such paper, by *Riadh Ben Messaoud, Sabine Loudcher Rabaséda, Omar Boussaid, and Rokia Missaoui* is entitled "*Enhanced Mining of Association Rules from Data Cubes*" and was presented by Riadh Ben Messaoud. The paper is based on the observation that although traditional research has spent significant effort to mine association rules from alphanumeric data, the multidimensional nature of such data was not taken into consideration. The paper presents a detailed taxonomy of the state of the art. Riadh in his presentation went on to explain how inter-dimensional rules can be extracted from data cubes, by exploiting parts of the cube each time. The cube is organized in context and analysis dimensions that allow the user to specify which part of the cube is actually interesting for him over which analysis dimensions. This formulation, combined with interestingness criteria allows the efficient computation of association rules. Riadh presented an algorithm and experimental results for this task.

The second paper in this session, by *Marc Plantevit*, *Anne Laurent*, and *Maguelonne Teisseire* is entitled "*HYPE: Mining Hierarchical Sequential Patterns*" and was presented by Marc Plantevit. The paper is based on the observation that no algorithms on the extraction of sequential patterns exploit hierarchical information, provided by OLAP modeling. The authors provide a rigorous treatment of an OLAP environment, specifically tailored to their end of mining hierarchical patterns. Specific emphasis is placed to the notion of support, since the hierarchical space can be decomposed in blocks, with different support counting for each block. Mark explained the experiments of the proposed algorithm and the improvements achieved over the current state-of-the-art on the topic.

The third paper of the session, by *Michael Mathioudakis*, *Dimitris Sacharidis* and *Timos Sellis*, is entitled "*A Study on Workload-aware Wavelet Synopses for Point and RangeSum Queries*" and was presented by Michael Mathioudakis. The paper is about an approximation approach and Michael started with a survey of approximation efforts, specifically, wavelets, to answer decision support, summary queries. Michael indicated that taking query frequencies into consideration significantly increases the complexity of assessing the error of state-of-the-art techniques for point and range-restricted summary queries. The tuning of the overall process was rigorously dealt with and results on an algorithm improving the state-of-the-art for range-sum queries were also presented.

In the final paper of this session, by *Igor Timko*, *Curtis E. Dyreson*, and *Torben Bach Pedersen*, entitled "*Pre-Aggregation with Probability Distributions*", Igor Timko presented an original approach towards the summarization of information about probabilistic data in the general case (and location-based data, in particular). The general problem addressed has to do with aggregating measures of probabilistic data, which is particular difficult due to the high number of alternative summarization groups involved. The approximation of the summaries is therefore necessary and the problem is the identification of the appropriate value intervals that will allow the summarization of data without loss of important information. The proposed pre-aggregation method deals with (a) the correct identification of a probabilistic description of the factual data along different levels of dimension hierarchies and (b) with the computation of error-free summations.

2.3 Data Warehouse Construction

The session on data warehouse construction involved papers reporting on case-studies and providing insights for real-world problems and topics for future research indicated by them. The first paper of the session, by *Christian Thomsen* and *Torben Bach Pedersen*, entitled "*Building a Web Warehouse for Accessibility Data*", discusses the case of real-world, web warehouse for the European Internet Accessibility Observatory (EIAO) project, a crawler that will evaluate the accessibility of several European web sites concerning people with disabilities. Christian Thomsen in his talk explained the motivation and organization of the warehouse. Parts of the design blueprints and the internal architecture of the warehouse were presented. Christian explained some of the benefits of the design choices made and reported on problems encountered, mainly due to the volatile structure of the warehouse's sources and the vast volume of collected data.

The second paper of the session, by *Matteo Golfarelli*, *Stefano Rizzi*, and *Andrea Proli* entitled "*Designing What-if Analysis: Towards a Methodology*" was presented by Matteo Golfarelli. Matteo introduced foresights for future research in the area of what-if analysis, which can be described as "a data-intensive simulation whose goal is to inspect the behavior of a complex system under some given hypotheses". Matteo described requirements and state-of-the-practice for tools involved in what-if analysis. Methodological sketches and real-world problems were also presented by the Matteo. The experiences of the authors with a large Italian company have driven the authors to conclude with the fact that concrete methodological approaches are necessary for the design of warehouses to support what-if-analyses, combined with rigorous and algorithmic results for the extension of OLAP tools with what-if operators. Finally, methods to allow the user express scenarios and knowledge efficiently was also highlighted as a topic of future research.

2.4 Design Issues in Data Warehousing

The design phase of a data warehouse has traditionally dominated the agenda of the research community, due to the complexity and the high risks involved in this part of the data warehouse lifecycle. DOLAP 2006 could not be an exception and four papers on the topic were presented.

In their paper, "*Heuristic Design of Property Maps*", *Ravi Darira*, *Karen C. Davis* and *Jennifer Grommon-Litton* discuss the efficient construction of Property Maps. The Property Map is a specialized multidimensional indexing technique that precomputes attribute expressions for selected data items and stores the results as bit strings. Property Maps effectively index multi-attribute queries and high cardinality attributes. Karen Davis in her talk explained the difficulties in the design of Property Maps, mainly due to the vast, exponential space to be explored. Then, she went on to describe a proposed heuristic algorithm that detects common areas in the involved attribute expressions and thus, reduces the number of solutions to consider while still producing a Property Map index with good performance. Karen also discussed implementation details and experimental results for the proposed approach.

The second paper of the session was authored by *Dimitris Skoutas* and *Alkis Simitsis*, with the title "*Designing ETL Processes Using Semantic Web Technologies*". In this paper, Semantic Web technologies are exploited to facilitate the process of selecting the relevant information from the available data sources and appropriately transforming it to populate the data warehouse. The approach is based on the idea of exploiting an ontology to annotate data source and data warehouse schemata, in order to formally and explicitly specify the semantics of their interrelationships. OWL is used as the formal language for the construction of the ontology, based on the initial conceptual model of the warehouse. The final design of the ETL workflow is then performed in a semi-automatic way.

In the next paper of the session, by *Wugang Xu*, *Dimitri Theodoratos*, and *Calisto Zuzarte*, entitled "*Computing Closest Common Subexpressions for View Selection Problems*", Dimitris Theodoratos presented an approach for dealing with the view selection (and materialization) problem. Given a set of input queries the objective is to determine which views should be materialized in order to answer the queries as efficiently as possible. Part of the problem involves the detection of common parts in the queries. Once common parts are detected, then they can be materialized and serve all the queries to which they belong. Naturally, query rewriting is also necessary, since queries need to be reformulated over the newly introduced materialized views. Dimitris presented a rigorous model for the problem and an algorithm for the detection of common subexpressions that improves the current state-of-the-art in the field.

Finally, the paper "*Towards a Logical Multidimensional Model for Spatial Data Warehousing and OLAP*" on spatial data warehouse by *Marcus Sampaio, André Sousa,*and *Cláudio Baptista*, presented by Cláudio Baptista, discusses a multidimensional meta-model for spatial data warehouses and spatial aggregations. The metamodel was used in a real-world case study through a tool, MapWarehouse. The case of spatial roll-ups was detailed and illustrated extensively; in fact, the visual representation of facts combined with aggregate information is indeed the gist of spatial warehousing. Experimental results for query optimization in Oracle environment were also presented.

3 Conclusions

DOLAP '06 brought new insights on the future research topics in the field. We recommend that the following hot topics be further researched in the data warehouse and OLAP community: exploration of business intelligence by utilizing active data warehousing technologies and advanced OLAP operations, novel query mechanisms (through approximate data structures or approximation techniques), techniques for the engineering of the data warehouse construction, and more powerful architectures to incorporate novel kinds of data and applications like, web, streaming, GIS, XML, or biomedical data, along with orthogonal considerations like security or data quality control.

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'06, the DOLAP Steering Committee and our sponsors for their support in organizing this workshop.