Report on the ACM Sixth International Workshop on Data Warehousing and OLAP (DOLAP 2003)

Stefano Rizzi
University of Bologna
Bologna, Italy
srizzi@deis.unibo.it

Il-Yeol Song
Drexel University
Philadelphia, USA
song@drexel.edu

1. Introduction

Data Warehouses are specialized databases for intelligent business applications. Data Warehouses have been rapidly spreading in the industrial world over the last decade due to their undeniable contribution to increasing the effectiveness and efficiency of the decision process within business and scientific domains. This wide diffusion was supported by remarkable research results on the one hand, and by the quick advancement of commercial tools on the other. However, several issues still need to be deeply investigated and well understood in order to achieve full maturity for the technologies employed in data warehousing systems.

The DOLAP (*ACM International Workshop on Data Warehousing and OLAP*) series of workshops, which started in 1998, has become a landmark for researchers and practitioners in the data warehousing field over the past years. DOLAP provides an exciting forum of discussions for innovative ideas and solutions in both basic and applied research. The sixth edition of DOLAP was held on November 7, 2003 in New Orleans, in conjunction with CIKM'03. It was a full-day workshop mainly aimed at reducing the gap between the research community and industry practitioners.

Though the workshop call for papers solicited regular and industry papers, only 2 industry papers were submitted out of 28 total submissions. This was not a surprise, since traditionally industry papers tend to be submitted to DMDW rather than to DOLAP. After a very careful review process, 12 high-quality papers were selected among the submissions; these papers were organized into four sessions, described in Section 2: OLAP; XML and architecture; Query processing; Maintenance and workload.

Unfortunately, the authors of three accepted papers could not make it to the workshop due to last-minute problems. On the other hand, this left room for a fruitful panel discussion, involving all workshop participants, whose main outcomes are summarized in Section 3.

2. Paper presentations

Historically, DOLAP has always been a reliable indicator to measure the current trends in data warehouse research. This year we witnessed the appearance of some non-traditional topics, such as workload and maintenance, OLAP applications to non-business domains, OLAP-XML, spatial data warehouses, and visualization. Conversely, some of the preferred topics in previous editions, such as conceptual modeling and design, virtually disappeared from the range of submissions. Noticeably, this was the first DOLAP where no paper on view materialization was presented.

In this section we summarize the main features of the paper presentation. The full proceedings of DOLAP 2003 are available at http://www.cis.drexel.edu/faculty/song/dolap03/dolapmenu.htm. The slides of the presentations are also available in the web site.

2.1 OLAP

This session started with Owen Kaser presenting a paper about attribute value reordering for efficient hybrid OLAP. The problem is that of computing the optimal ordering for attribute values, where the chosen ordering determines dense and sparse subcubes thus affecting the physical storage of data and, ultimately, querying efficiency. It was shown that the problem is NP-hard even for small subcubes, and some effective heuristics were proposed and experimentally evaluated.

Again in the field of physical storage for OLAP, Yannis Sismanis described a work on hierarchical dwarfs for the rollup cube. Dwarf is a compressed structure for storing and indexing data cubes; in this work, dwarf is extended by incorporating cubes with hierarchical dimensions in order to efficiently support aggregate queries on the different levels of hierarchies.

In the field of OLAP visualization, Andreas Maniatis showed how the Cube Presentation Model proposed in earlier work can be mapped on the table lens, an advanced visualization technique suitable for cross-tab reports. Then, he provided an algorithm that proactively supports the user in exploring an OLAP report.

2.2 XML and architecture

XML has become a very hot topic over the last couple of years; this session focused on its impact on data warehousing techniques.

The first paper, presented by Torben Bach Pedersen, identified a number of potential problems in OLAP-XML federations and discussed some techniques to handle these problems aimed at increasing the system robustness. Specifically, techniques for managing changes in XML data sources and for increasing their reliability by locating alternative sources were proposed. The work was carried out in collaboration with an OLAP vendor, and led to extending and improving an industrial product.

The second paper, presented by Wolfgang Hümmer, proposed XCube: a family of XML templates used to exchange data cubes over a network. Three different use cases are considered: downloading cubes from a web server, querying cubes residing on a web server, and creating cubes on a web server. The main characteristic on XCube is modularity, which enables schema and data to be separately transmitted.

2.3 Query processing

Since the very first editions, query processing has always been a popular topic for DOLAP. This year, we had two presentations related to query processing.

In the first talk, Fang Yan Rao from IBM Research Laboratory described an approach to include spatial hierarchies in star schemas; besides, she proposed and evaluated a technique based on the Generalized Index Searching Tree to improve performances in spatial data warehouses.

The session was closed by Albert Abelló's presentation, which discussed how to implement operations to navigate semantic star schemas. After proposing the translation of an algebraic set of cube operations to SQL in relational environments, he focused on drill-across queries to show how the subject of analysis can be changed even if dimensions do not exactly coincide, by exploiting semantic relationships.

2.4 Maintenance and workload

The last session featured two presentations concerning not-so-frequently visited topics in data warehousing: maintenance and workloads.

The first presentation, by Henrik Engström, discussed a heuristic approach for selecting maintenance policies in data warehouses with multiple heterogeneous sources. A maintenance policy describes how to detect and propagate changes from distributed sources to the warehouse. Different policies are possible; a set of heuristics to guide in policy selection are proposed and evaluated by means of a prototype tool.

In the second talk, Matteo Golfarelli proposed an original set of indicators to profile an OLAP workload. Then, he described a profile-based algorithm to generate workloads for benchmarking and showed the results of a set of tests aimed at demonstrating the effectiveness of profiling with reference to optimization techniques like view materialization and indexing.

3. Panel discussion

The panel discussion was aimed at summarizing the state-of-the-art of research in data warehousing and OLAP, at identifying the main shortcomings in the field, and at conjecturing some directions for future research.

We started by observing that research can be considered to be mature by now, since most problems were at least attacked and consolidated solutions for all "easy" problems are available. On the industrial side, business activities are fully in action as demonstrated by the fact that about 80% of US Top-1000 companies are currently involved in data warehousing projects. On the other hand, we recognized that data warehousing has finally evolved into the mainstreams of business intelligence, a wider discipline incorporating CRM, data mining, knowledge management and other interdisciplinary issues such as, for instance, those related to devising effective key performance indicators to monitor business activities in directional cockpits.

The discussion gave us a chance to repeat that unfortunately, as already stated at DMDW 2003 in Berlin, there still is a significant gap between research and industries. Few of the most advanced research results have been implemented into commercial systems so far (for instance, most tools currently implement only basic approaches to query rewriting). On the other hand, vendors feel that some key practical issues were overlooked by research (for instance, little attention has been paid to managing the versioning of data warehouse schemas and allowing them to be related in a single query). Besides, we registered a dramatic lack of standards imposed by the market or agreed upon by the academic world, concerning for instance conceptual and logical models, design methodologies, and cube operators. We concluded that, though obviously there is no trivial treatment for this disease, we should strive to rethink the role of research in order to find a closer connection with industrial needs as well as users' needs. It was agreed that researchers should be encouraged to get involved in real data warehousing projects, and to actively and critically use commercial tools. On the other hand, major efforts should be devoted to stimulate people from industries to participate in DOLAP, especially now that the DMDW series, which traditionally attracted several good industrial papers, has ceased.

4. Conclusions

The workshop participants concluded the following three points.

(1) Data-warehousing & OLAP research is getting mature, but there are still gaps between academic research and real-world problems. Attendants agreed that focused workshops such as DOLAP should be continuously held to further explore these research topics and share research results.

(2) DOLAP needs to include more participants from industry. This will bring more interesting interactions between research and industrial communities, possibly bringing out real-world oriented research topics, and help reduce the gap between the two communities. One way to support this idea is to offer a separate industry track that will not compete with the research papers during the review process.

(3) DOLAP needs to expand its scope to include recent developments in related technologies such as business intelligence, web services, and knowledge management. In addition, we expect that the scope of data warehousing and OLAP technology will be applied to non-traditional applications such as cyber security, scientific and engineering applications. For example, data in the areas of medical informatics and bioinformatics are rapidly increasing. It would be interesting to watch how data warehousing & OLAP technology can contribute to those new application areas. We encourage researchers and practitioners in those areas to submit papers to DOLAP 2004. It is expected that DOLAP will go through changes in terms of its scope and even possibly its name to reflect those new developments.

Acknowledgements

DOLAP 2003 was sponsored by ACM's Special Interest Groups on Information Retrieval and Management Information Systems. We would like to thank the program committee of DOLAP 2003 for the precious reviewing effort, the organizing committee of CIKM 2003 for their support for registration, and the DOLAP Steering Committee for their unrelenting support.