Lecture Notes in Computer Science 2712
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
*Berlin*
*Heidelberg*
*New York*
*Hong Kong*
*London*
*Milan*
*Paris*
*Tokyo*

Anne James Brian Lings
Muhammad Younas (Eds.)

# New Horizons in Information Management

20th British National Conference on Databases, BNCOD 20
Coventry, UK, July 15-17, 2003
Proceedings

Springer

Volume Editors

Anne James
Muhammad Younas
Coventry University, School of Mathematical and Information Sciences
Priory Street, Coventry CV1 5FB, UK
E-mail: {A.James,M.Younas}@coventry.ac.uk

Brian Lings
University of Exeter, Department of Computer Science
Prince of Wales Road, Exeter EX4 4PT, UK
E-mail: B.J.Lings@exeter.ac.uk

Cataloging-in-Publication Data applied for

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;

detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): H.2-4

ISSN 0302-9743
ISBN 3-540-40536-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by PTP-Berlin GmbH
Printed on acid-free paper SPIN: 10928776 06/3142 5 4 3 2 1 0

# Preface

The British National Conference on Databases (BNCOD) was established in 1980 as a forum for research into the theory and practice of databases. The 20th in the series (BNCOD 20) was held at Coventry University in July 2003. This volume contains the BNCOD 20 proceedings. It was a delight to welcome researchers from all over the world to BNCOD 20. A strong response to our call for research papers together with the thoughtful work of the programme committee led to an excellent technical programme covering many fascinating challenges facing the database world today.

The theme of BNCOD 20 was "New Horizons in Information Management". Over the last decades database technology has become embedded in most of the information systems we use in commerce and industry and has been proven to be an essential tool in information management. Advances in database technology have enabled new methods of working with information. Similarly, new requirements of information systems have led to extensions of database technology. Novel application areas demand further development and integration of database theory within emerging fields. BNCOD 20 called for papers on new directions in information management and how database techniques are being adapted to support these. The global technological infrastructure is considered to be particularly pertinent. Thus areas such as the Semantic Web, the Hidden Web, information systems integration, information retrieval, co-operative working and Web-based agents, as they relate to databases, were of considerable interest.

BNCOD 20 had great pleasure in receiving two eminent invited speakers: Prof. Malcolm Atkinson from the Universities of Glasgow and Edinburgh, Scotland, UK and Director of the National e-Science Centre, Scotland, UK and Prof. Hector Garcia-Molina from Stanford University, USA. Both considered the challenges and opportunities afforded by the new technological environment.

Malcolm Atkinson has a long track record of contributions to research in large and long-lived systems. He is currently working on GridNet, a project to establish a UK Network of Excellence in Grid Computing and e-Science. He is also working on the auto-optimization of highly scalable index frameworks for large collections of reference data and developing techniques and software to support management and monitoring of user actions in distributed systems. In this invited presentation Malcolm explores the Grid and e-Science. He raises the question of how DBMS will fit into the "global computing machine" and considers the extent to which current database solutions fit the new Grid operation model.

Hector Garcia-Molina is the Leonard Bosack and Sandra Lerner Professor in the Departments of Computer Science and Electrical Engineering at Stanford University, Stanford, USA. He is the current chairman of the Computer Science Department and was the recipient of the 1999 ACM SIGMOD Innovations Award. His research interests include distributed computing systems, digital libraries and database systems. Hector is currently involved in developing improved methods for searching the Web and also techniques for managing peer-to-peer networks. In his presentation he addressed the area of Web crawling. Web crawlers can consume significant resources. Hector discusses how efficient crawlers can be built and how the Hidden Web can be accessed.

The refereed papers are presented in five sessions. The first refereed paper session is on XML and Semistructured Data and contains two full and two short papers. The first paper by Vincent et al. defines multivalued dependencies in XML (XMVDs). It shows that, for a very general class of mappings from relations to XML, a relation satisfies an MVD if and only if the corresponding XML document satisfies the corresponding XMVD. Thus they consider their definition of XMVD in XML documents to be a natural extension of the definition of MVD in relations. Moon et al. investigate the problem of processing an XML path expression using an XML cache maintained as materialized views. They propose algorithms to rewrite the given XML path expression using its relevant materialized view, and also provide implementation details of their approach. The paper by Pandrangi et al. describes WebVigil, a system for detecting changes in Web pages based on user profiles. Although this approach is discussed in the context of HTML and XML it can be generalized to other technologies. The session ends with two short papers, which are related to ontologies. The first paper by Bi et al. describes a query paradigm based on ontologies, aggregate table-driven querying, and expansion of QBE. The authors claim two novel features: visually specifying aggregate table queries and table layout in a single process and providing users with an ontology guide in composing complex analysis tasks as queries. The final short paper by Volz considers the requirements for a new level of data independence for ontology-based applications. For example, customization for other agents may be required. It proposes a solution based on the idea of integrative external ontologies.

The second refereed paper session is on Performance in Searching and Mining and contains three full and one short paper. The first paper by Garcia-Arellano et al. evaluates the relative performance of the IQ-tree and the A-tree in similarity search in high-dimensional data spaces. They introduce the Clustered IQ-tree, which is an indexing strategy that combines the best features of the IQ tree and the A-tree leading to a seemingly better and more stable performance over different types of data set. The paper by Mishra et al. concentrates on the K-way join approach, a technique for mining data irrespective of stored format. The authors look at various optimization methods for the K-way join and evaluate them. This work aims at feeding the results into an optimizer for data mining. The paper by Yiannis et al. explores the effects of compression on the cost of external sorting. Whilst compression can often be useful, on-the-fly compression can be slow and some compression techniques do not allow random access to individual records. Yiannis et al. look at these issues for various techniques and develop improved solutions. They show that incorporation of compression can significantly accelerate query processing. The final paper of the session is a short paper by Srikumar et al. In the paper the authors present MaxDomino, an algorithm for mining maximal frequent sets using the novel concept of the dominancy factor of a transaction.

Transformation, integration and extension were the topics of the third refereed paper session. This session contained three full and two short papers. Engstrom et al. consider maintenance policies for externally materialized multi-source views. They consider various methods and show that in all situations it is more efficient to use auxiliary views than policies which require consistency-preserving algorithms. Tong considers transformation optimization techniques within the automed database integration system. A new representation of schema transformations is presented.

These are claimed to facilitate the verification of well-formedness and the optimization of the transformation sequences. The paper by Scallehn et al. raises the issue of dealing with discrepancies in data integration. The authors present similarity-based variants of grouping and join operations as a solution to this problem of attributes that are similar but not equal. The first of the two short papers in this session is by Green et al. They describe ProSQL, a prototyping tool to support the development of extensions to SQL. The system was developed by building a wrapper around an existing DBMS and providing a collection of interfaces through which a designer can define extensions to the basic relational database. The final short paper by Al-Mourad et al. addresses the problem of integrating object-oriented schemas with multiple behaviour requirements. The Multiple Views supporting the Multiple Behaviours System (MVMBS) is described.

The fourth refereed paper session was on events and transactions and included two full and two short papers. Hinze tackles the problem of rapid notification of composite events. Currently the detection of composite events requires a second filtering step after the identification of the primitive components. Hinze proposes a single-step method for the filtering of composite events, and presents results which show improvement in performance for event filtering. Ray investigates an interesting issue of multi-level secure (MLS) active database systems by defining MLS rules and assigning them security levels. Ray also determines the impact of MLS rules on the execution models of existing active database systems. The short paper by Howard et al. describes a Compliant Systems Architecture (CSA) and shows how it can deliver flexibility within a two-phase commit protocol of distributed transactions. CSA aims at providing strict separation of policy and mechanism. Lim et al. present a new concurrent $B^{link}$-tree algorithm that provides a concurrent tree restructuring mechanism for handling underflow nodes as well as overflow nodes.

Two short papers are delivered in the final refereed paper session, which is on Personalisation and the Web. Dempster et al. discuss a framework for personalisation and an initial prototype toolkit. Cooper et al. propose an approach to information extraction from e-mail text, which involves creating sentence structures from metadata, pattern-matching, and generating update statements.

Once again BNCOD yielded an excellent range of papers. This was through the industry and interest of our international research community and this is much appreciated. The pivotal role of databases in information systems continues to interest, challenge and provide opportunities for the development of new and improved systems.

## Acknowledgements

We would like to thank the programme committee for their excellent work in reviewing and providing comments on the many papers submitted to the conference. Once again their dedication and commitment helped to produce another inspiring and technical programme of the high standard expected of the BNCOD series. Thanks go also to Alex Gray for inviting us to organize BNCOD20 and for providing useful advice and enthusiasm throughout. Thanks also to Mary Garvey and Mike Jackson for help with the organization and sharing of ideas. The administrative support of Serena Morgan and Rachel Carter was most appreciated and likewise the help provided at the conference by Yih-Ling Hedley, Rahat Iqbal and Mofed Salem.

April 2003 Anne James, Brian Lings, Muhammad Younas

# Conference Committees

## Programme Committee

| | |
|---|---|
| Brian Lings (Chair) | University of Exeter |
| David Bell | University of Ulster |
| Peter Buneman | University of Edinburgh |
| Barry Eaglestone | University of Sheffield |
| Suzanne Embury | University of Manchester |
| Alex Gray | University of Wales, Cardiff |
| Peter Gray | University of Aberdeen |
| Mike Jackson | University of Wolverhampton |
| Anne James | Coventry University |
| Keith Jeffery | CLRC Rutherford Appleton |
| Jessie Kennedy | Napier University |
| Nigel Martin | Birkbeck College, University of London |
| Peter McBrien | Imperial College, University of London |
| Ken Moody | University of Cambridge |
| Werner Nutt, | Heriot-Watt University |
| Norman Paton | University of Manchester |
| Alexandra Poulovassilis | Birkbeck College, University of London |
| Brian Read | London Metropolitan University |
| Howard Williams | Heriot-Watt University |
| Muhammad Younas | Coventry University |

## Organizing Committee

| | |
|---|---|
| Anne James (Chair) | Coventry University |
| Mary Garvey | University of Wolverhampton |
| Alex Gray | University of Wales, Cardiff |
| Mike Jackson (Prizes Chair) | University of Wolverhampton |
| Muhammad Younas | Coventry University |

## Steering Committee

| | |
|---|---|
| Alex Gray (Chair) | University of Wales, Cardiff |
| Carole Goble | University of Manchester |
| Barry Eaglestone | University of Sheffield |
| Keith Jeffery | CLRC Rutherford Appleton |
| Roger Johnson | Birkbeck College, University of London |
| Brian Lings | University of Exeter |

# Table of Contents

## Invited Papers

## XML and Semi-structured Data

## Performance in Searching and Mining

## Transformation, Integration, and Extension

## Events and Transactions

## Personalisation and the Web