Lei Chen
Chengfei Liu
Qing Liu
Ke Deng (Eds.)

LNCS 5667

Database Systems for Advanced Applications

DASFAA 2009 International Workshops: BenchmarX, MCIS, WDPP, PPDA, MBC, PhD Brisbane, Australia, April 2009

Springer

Lecture Notes in Computer Science 5667

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Lei Chen Chengfei Liu
Qing Liu Ke Deng (Eds.)

Database Systems for Advanced Applications

DASFAA 2009 International Workshops:
BenchmarX, MCIS, WDPP, PPDA, MBC, PhD
Brisbane, Australia, April 20-23, 2009

Springer

Volume Editors

Lei Chen
Hong Kong University of Science and Technology
E-mail: leichen@cse.ust.hk
Chengfei Liu
Swinburne University of Technology, Melbourne, Australia
E-mail: cliu@swin.edu.au
Qing Liu
CSIRO, Castray Esplanade, Hobart, TAS 7000, Australia
E-mail: q.liu@csiro.au
Ke Deng
The University of Queensland, Brisbane, QLD 4072, Australia
E-mail: dengke@itee.uq.edu.au

Library of Congress Control Number: 2009933477

CR Subject Classification (1998): H.2, H.3, H.4, H.5, J.1, H.2.4, H.3.4, K.6.5, I.7

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-04204-X Springer Berlin Heidelberg New York
ISBN-13 978-3-642-04204-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12743681 06/3180 5 4 3 2 1 0

Preface

DASFAA is an annual international database conference, located in the Asia-Pacific region, which showcases state-of-the-art R & D activities in database systems and their applications. It provides a forum for technical presentations and discussions among database researchers, developers and users from academia, business and industry. DASFAA 2009, the 14th in the series, was held during April 20-23, 2009 in Brisbane, Australia.

In this year, we carefully selected six workshops, each focusing on specific research issues that contribute to the main themes of the DASFAA conference. This volume contains the final versions of papers accepted for these six workshops that were held in conjunction with DASFAA 2009. They are:

- First International Workshop on Benchmarking of XML and Semantic Web Applications (BenchmarX 2009)
- Second International Workshop on Managing Data Quality in Collaborative Information Systems (MCIS 2009)
- First International Workshop on Data and Process Provenance (WDPP 2009)
- First International Workshop on Privacy-Preserving Data Analysis (PPDA 2009)
- First International Workshop on Mobile Business Collaboration (MBC 2009)
- DASFAA 2009 PhD Workshop

All the workshops were selected via a public call-for-proposals process. The workshop organizers put a tremendous amount of effort into soliciting and selecting papers with a balance of high quality, new ideas and new applications. We asked all workshops to follow a rigid paper selection process, including the procedure to ensure that any Program Committee members are excluded from the paper review process of any paper they are involved with. A requirement about the overall paper acceptance rate of no more than 50% was also imposed on all the workshops.

The conference and the workshops received generous financial support from The University of Melbourne, The University of New South Wales, The University of Sydney, The University of Queensland, National ICT Australia (NICTA), Australian Research Council (ARC) Research Network in Enterprise Information Infrastructure (EII), ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), and ARC Research Network for a Secure Australia. We also received extensive help and logistic support from the DASFAA Steering Committee, The University of Queensland, Tokyo Institute of Technology and the Conference Management Toolkit Support Team at Microsoft.

We are very grateful to Xiaofang Zhou, Haruo Yokota, Qing Liu, Ke Deng, Shazia Sadiq, Gabriel Fung, Kathleen Williamson and many other people for

their effort in supporting the workshop organization. We would like to take this opportunity to thank all workshop organizers and Program Committee members for their effort to put together the workshop program of DASFAA 2009.

April 2009

Lei Chen
Chengfei Liu

DASFAA 2009 Workshop Organization

Workshop Chairs

Lei Chen	Hong Kong University of Science and Technology, China
Chengfei Liu	Swinburne University of Technology, Australia

Publication Chairs

Qing Liu	CSIRO, Australia
Ke Deng	University of Queensland, Australia

BenchmarX 2009 Workshop

Program Committee Chairs

Michal Kratky	Technical University of Ostrava, Czech Republic
Irena Mlynkova	Charles University in Prague, Czech Republic
Eric Pardede	La Trobe University, Bundoora, Australia

Program Committee

Radim Baca	Technical University of Ostrava, Czech Republic
Martine Collard	INRIA Sophia Antipolis, France
Jiri Dokulil	Charles University in Prague, Czech Republic
Peter Gursky	Pavol Jozef Safarik University in Kosice, Slovakia
Tomas Horvath	Pavol Jozef Safarik University in Kosice, Slovakia
Jana Katreniakova	Comenius University in Bratislava, Slovakia
Markus Kirchberg	Institute for Infocomm Research, A*STAR, Singapore
Agnes Koschmider	Institute AIFB, Universität Karlsruhe, Germany
Michal Kratky	Technical University of Ostrava, Czech Republic
Sebastian Link	Victoria University of Wellington, New Zealand

Pavel Loupal	Czech Technical University in Prague, Czech Republic
Mary Ann Malloy	The MITRE Corporation, USA
Marco Mevius	Institute AIFB, Universität Karlsruhe, Germany
Irena Mlynkova	Charles University in Prague, Czech Republic
Martin Necasky	Charles University in Prague, Czech Republic
Alexander Paar	Universität Karlsruhe, Germany
Incheon Paik	The University of Aizu, Japan
Eric Pardede	La Trobe University, Bundoora, Australia
Jorge Perez	Pontificia Universidad Catolica, Chile
Dmitry Shaporenkov	University of Saint Petersburg, Russia
Michal Valenta	Czech Technical University in Prague, Czech Republic

MCIS 2009 and WDPP 2009 Workshops

Program Committee Chairs

MCIS

Shazia Sadiq	The University of Queensland, Australia
Ke Deng	The University of Queensland, Australia
Xiaofang Zhou	The University of Queensland, Australia
Xiaochun Yang	Northeastern University, China

WDPP

Walid G. Aref	Purdue University, USA
Alex Delis	University of Athens, Greece
Qing Liu	CSIRO ICT Centre, Australia

Publicity Chair *(*WDPP*)*

Kai Xu	CSIRO ICT Centre, Australia

Program Committee

MCIS

Yi Chen	Arizona State University, USA
Markus Helfert	Dublin City University, UK
Ruoming Jin	Kent State University, USA
Chen Li	UC Irvine, USA
Jiaheng Lu	Renmin University, China
Graeme Shanks	Monash University, Australia
Can Turker	FGCZ Zurich, Switzerland
Haixun Wang	IBM, USA
Xuemin Lin	UNSW, Australia

WDPP

Mohamed S. Abougabal	University of Alexandria, Egypt
Ilkay Altintas	San Diego Supercomputer Centre, USA
Athman Bouguettaya	CSIRO, Australia
Susan B. Davidson	University of Pennsylvania, USA
Antonios Deligiannakis	Technical University of Crete, Greece
Juliana Freire	University of Utab, USA
James Frew	University of California, Santa Barbara, USA
Paul Groth	University of Southern California, USA
Georgia Koutrika	Stanford University, USA
Bertram Ludascher	University of California, Davis, USA
Simon McBride	The Australian E-Health Research Centre, Australia
Simon Miles	King's College London, UK
Brahim Medjahed	University of Michigan, Dearborn, USA
Khaled Nagi	Alexandria University, Egypt
Anne Ngu	Texas State University, San Marcos, USA
Mourad Ouzzani	Purdue University, USA
Thomas Risse	L3S Lab, Germany
Satya S. Sahoo	Kno.e.sis Center, Wright State University, USA
Zahir Tari	RMIT, Australia
Qi Yu	Rochester Institute of Technology, USA
Jun Zhao	University of Oxford, UK

PPDA 2009 Workshop

Raymond Chi-Wing Wong	The Hong Kong University of Science and Technology, China
Ada Wai-Chee Fu	The Chinese University of Hong Kong, China

Program Committee

Claudio Bettini	University of Milan, Italy
Chris Clifton	Purdue University, USA
Claudia Diaz	K.U.Leuven, Belgium
Josep Domingo-Ferrer	Rovira i Virgili University, Spain
Elena Ferrari	University of Insubria, Italy
Sara Foresti	University of Milan, Italy
Benjamin C.M. Fung	Concordia University, Canada
Christopher Andrew Leckie	The University of Melbourne, Australia
Jiuyong Li	University of South Australia, Australia
Jun-Lin Lin	Yuan Ze University, Taiwan

Kun Liu	IBM Almaden Research Center, USA
Ashwin Machanavajjhala	Cornell University, USA
Bradley Malin	Vanderbilt University, USA
Nikos Mamoulis	Hong Kong University, China
Wee Keong Ng	Nanyang Technological University, Singapore
Jian Pei	Simon Fraser University, Canada
Yucel Saygin	Sabanci University, Turkey
Jianhua Shao	Cardiff University, UK
Yufei Tao	The Chinese University of Hong Kong, China
Vicenc Torra	Universitat Autonoma de Barcelona, Spain
Carmela Troncoso	K.U. Leuven, Belgium
Hua Wang	University of Southern Queensland, Australia
Ke Wang	Simon Fraser University, Canada
Sean Wang	University of Vermont, USA
Duminda Wijesekera	George Mason University, USA
Xintao Wu	University of North Carolina at Charlotte, USA
Jeffrey Yu	The Chinese University of Hong Kong, China
Philip S. Yu	University of Illinois at Chicago, USA

MBC 2009 Workshop

General Chairs

Qing Li	City University of Hong Kong, China
Hua Hu	Zhejiang Gongshang University, China

Program Committee Chairs

Dickson K.W. Chiu	Dickson Computer Systems, Hong Kong, China
Yi Zhuang	Zhejiang Gongshang University, China

Program Committee

Patrick C.K. Hung	University of Ontario Institute of Technology, Canada
Samuel P.M. Choi	The Open University of Hong Kong, China
Eleanna Kafeza	Athens University of Economics and Commerce, Greece
Baihua Zheng	Singapore Management University, Singapore
Edward Hung	Hong Kong Polytechnic University, China
Ho-fung Leung	Chinese University of Hong, China
Zakaria Maamar	Zayed University, UAE
Stefan Voss	University of Hamburg, Germany
Cuiping Li	Renmin University, China

Chi-hung Chi	National Tsing Hua University, Taiwan, China
Stephen Yang	National Central University, Taiwan, China
Ibrahim Kushchu	Mobile Government Consortium International, UK
Haiyang Hu	Zhejiang Gongshang University, China
Huiye Ma	CWI, The Netherlands
Pirkko Walden	Abo Akademi University, Finland
Raymond Wong	National ICT, Australia
Lidan Shou	Zhejiang University, China
Matti Rossi	Helsinki School of Economics, Finland
Achim Karduck	Furtwangen University, Germany

DASFAA 2009 PhD Workshop

Program Committee Chairs

Wei Wang	University of New South Wales, Australia
Baihua Zheng	Singapore Management University, Singapore

Program Committee

Bin Cui	Peking University, China
Jianlin Feng	Zhongshan University, China
Haifeng Jiang	Google, USA
Takahiro Hara	Osaka University, Japan
Wang-chien Lee	Pennsylvania State University, USA
Jiaheng Lu	Renming University, China
Lidan Shou	Zhejiang University, China
Bill Shui	NICTA, Australia
Xueyan Tang	Nanyang Technological Universtiy, Singapore
Jianliang Xu	Hong Kong Baptist University, China

Table of Contents

First International Workshop on Benchmarking of XML and Semantic Web Applications (BenchmarX'09))

Second International Workshop on Managing Data Quality in Collaborative Information Systems and First International Workshop on Data and Process Provenance (MCIS'09 & WDPP'09)

First International Workshop on Privacy-Preserving Data Analysis (PPDA'09)

First International Workshop on Mobile Business Collaboration (MBC'09)

DASFAA 2009 PhD Workshop

First International Workshop on Benchmarking of XML and Semantic Web Applications (BenchmarX'09)

Workshop Organizers' Message

Michal Kratky [1], Irena Mlynkova [2], and Eric Pardede [3]

[1] Technical University of Ostrava, Czech Republic
[2] Charles University in Prague, Czech Republic
[3] La Trobe University, Bundoora, Australia

The 1st International Workshop on Benchmarking of XML and Semantic Web Applications (BenchmarX'09) was held on April 20, 2009 at the University of Queensland in Brisbane, Australia in conjunction with the 14th International Conference on Database Systems for Advanced Applications (DASFAA'09). It was organized by Jiri Dokulil, Irena Mlynkova and Martin Necasky from the Department of Software Engineering of the Charles University in Prague, Czech Republic.

The main motivation of the workshop was based on the observation that even though XML and semantic data processing is the main topic of many conferences around the world, the communities dealing with XML and semantic data benchmarking and related issues are still scattered. Moreover, although benchmarking is one of the key aspects of improvements of data processing, the majority of researches naturally concentrate on proposing new approaches, while benchmarking is often neglected. Therefore, the aim of BenchmarX was and is to bring the benchmarking research community together and to provide an opportunity to deal with this topic more thoroughly.

The program committee of the workshop consisted of 21 researchers and specialists representing 15 universities and institutions from 11 different countries. To ensure high objectiveness of the paper selection process 3 PC chairs from different institutions were selected, in particular Michal Kratky, Irena Mlynkova and Eric Pardede. Each of the submitted papers for BenchmarX'09 was reviewed by 3 PC members for its technical merit, originality, significance, and relevance to the workshop. Finally, the PC chairs decided to accept 40% of the submitted papers.

The final program of the workshop consisted of an invited talk and 2 sessions involving the accepted papers. The invitation was kindly accepted by Stephane Bressan from the National University of Singapore, one of the authors of the XOO7 benchmark and an expert in various aspects of data management.

Last but not least, let us mention that BenchmarX'09 would not be possible without the support of our sponsors. In particular it was partially supported by the Grant Agency of the Czech Republic, projects of GACR number 201/09/0990 and 201/09/P364.

After the successful first year providing many interesting ideas and research problems, we believe that BenchmarX will become a traditional annual meeting opportunity for the whole benchmarking community.

L. Chen et al. (Eds.): DASFAA 2009 Workshops, LNCS 5667, p. 3, 2009.

Current Approaches to XML Benchmarking
(Invited Talk)

Stéphane Bressan

School of Computing
National University of Singapore
steph@nus.edu.sg

Abstract. XML benchmarking is as versatile an issue as numerous and diverse are the potential applications of XML. It is however not yet clear which of these anticipated applications will be prevalent and which of their features and components will have such performance requirements that necessitate benchmarking.

The performance evaluation of XML-based systems, tools and techniques can either use benchmarks that consist of a predefined data set and workload or it can use a data set with an ad hoc workload. In both cases the data set can be real or synthetic. XML data generators such as Toxgene and Alphawork can generate XML documents whose characteristics, such as depth, breadth and various distributions, are controlled. It is also expected that benchmarks provide data generator with a fair amount of control of the size and shape of the data, if the data is synthesized, or offer a suite of data subsets of varying size and shape, if the data is real. Application level evaluation emphasizes the representativeness of the data set ad workload in terms of typical applications while micro-level evaluation focuses on elementary and individual technical features.

The dual view of XML, data view and document view, is reflected in its benchmarks. There exist several well established benchmarks for XML data management systems that can be used for the evaluation of the performance of query processing. The main application level benchmarks in this category are XOO7, XMach1, XMark, and XBench while The Michigan Benchmark is a micro-benchmark. For the evaluation of XML information retrieval the prevalent benchmark is the series of INEX corpora and topics. However, in practice, whether for the evaluation of XML data management techniques or for the evaluation of XML-retrieval techniques, researchers seem to favor real or synthetic data sets with ad hoc workloads when needed. The university of Washington repository gathers links to a variety of XML data sets. Noticeably most of these data sets are small. The largest is 603MB. Popular data sets like Mundial or the Baseball Boxscore XML are much smaller. The Database and Logic Programming Bibliography XML data set, also used by many scientists, is around 500MB. All of these data sets are generally relatively structured and quite shallow thus not necessarily conveying the expected challenges associated with the semi-structure nature of XML.

If the application level data sets and workloads are not satisfactory, It may well be the case that XML as a language used to structure and manage content has

L. Chen et al. (Eds.): DASFAA 2009 Workshops, LNCS 5667, pp. 4–5, 2009.

not yet matured. We must ask ourselves the question as to what is there really to benchmark. As of today, XML data are most commonly produced by office suites and software development kits. Office suites supporting Office Open XML and in Open Document Format are or will soon become the principal producers of XML. Yet in these environments XML is principally used to represent formatting instructions. Similarly, the widespread adoption of Web service standards in software development frameworks and kits (in the .Net framework, for instance) also contributes to the creation of large amounts of XML data. Again here XML is primarily used to represent formats (e.g. SOAP messages).

Although both XML-based document standards and Web service standards have intrinsic provision for XML content and have been designed to enable the management of content in XML, few users have yet the tools, the wants and the culture to manage their data in XML. Consequently, at least for now, it seems that these huge amounts of XML data created in the background of authoring and programming activities need neither be queried nor searched but rather only need to be processed by the office suites and compilers. The emphasis is still on format rather than content structuring and management. Of course, it is hoped by proponent of XML as a format for content that the XML-ization of formats will facilitate the XML-zation of the content.

With XML-based protocols and formats, XML as a "standards' standard" (as there are compiler compilers) has been most successful at the lower layers of information management. The efforts for content organization and management, on the other hand, do not seem to have been as pervasive and prolific (in terms of the amount of XML data produced and used). For instance, the volume of data in the much talked about business XML standards (Rosettanet or Universal Business Language, for instance) is still difficult to measure and may not be or become significant. In this presentation we critically review the existing approaches to benchmarking of XML-based systems and applications. We try to analyze the trends in the usage of XML and in order to determine the needs and requirements for the successful design, development and adoption of benchmarks.

CV: Stéphane Bressan is Associate Professor in the Computer Science department of the School of Computing (SoC) at the National University of Singapore (NUS). He joined the National University of Singapore in 1998. He is also adjunct Associate Professor at Malaysia University of Science and Technology (MUST) since 2004. He obtained his PhD in Computer Science from the University of Lille, France, in 1992. Stéphane was research scientist at the European Computer-industry Research Centre (ECRC), Munich, Germany, and at the Sloan School of Management of the Massachusetts Institute of Technology (MIT), Cambridge, USA. Stéphane's research is concerned with the management and integration of multi-modal and multimedia information from distributed, heterogeneous, and autonomous sources. He is author and co-author of more than 100 papers. He is co-author of the XOO7 benchmark for XML data management systems. Stéphane is member of the XML working group of the Singapore Information Technology Standards Committee (ITSC) and advisory member of the executive committee of the XML user group of Singapore (XMLone).

TJDewey – On the Efficient Path Labeling Scheme Holistic Approach*

Radim Bača and Michal Krátký

Department of Computer Science, Technical University of Ostrava
Czech Republic
{radim.baca,michal.kratky}@vsb.cz

Abstract. In recent years, many approaches to XML twig pattern searching have been developed. Holistic approaches are particularly significant in that they provide a theoretical model for optimal processing of some query classes and have very low main memory complexity. Holistic algorithms can be incorporated into XQuery algebra as a twig query pattern operator.

We can find two types of labeling schemes used by indexing methods: element and path labeling schemes. The path labeling scheme is a labeling scheme where we can extract all the ancestors labels from a node label. In the TJFast method, authors have introduced an application of the path labeling scheme (Extended Dewey) in the case of holistic methods. In our paper, we depict some improvements of this method that lead to a better scalability of the TJFast algorithm. We introduce the TJDewey algorithm which combines the TJFast algorithm with the DataGuide summary tree. The path labeling schemes have better update features and our article shows that the utilization of a path labeling scheme can have comparable or even better query processing parameters compared to other element labeling scheme approaches.

Keywords: XML, twig pattern query, holistic algorithms, path labeling scheme, TJFast.

1 Introduction

XML (*Extensible Mark-up Language*) [20] has recently been embraced as a new approach to data modeling. A *well-formed* XML document or a set of documents is an XML database. Implementation of a system enabling storage and querying of XML documents efficiently (the so-called *native XML databases*) requires an efficient approach to indexing the XML document structure.

Existing approaches to an XML document structure indexing use some kind of labeling scheme [22,6,19,1,17]. The labeling scheme associates every element of an XML document with a unique label, which allows us to determine the basic relationship between elements. We recognize two types of labeling schemes: (1) *element labeling scheme* (e.g., containment labeling scheme [22] or Dietz's labeling scheme [6]), and (2) *path labeling scheme* (e.g., Dewey order [19] or OrdPath [17]). By the term 'path

* Work is partially supported by Grants of GACR No. 201/09/0990.

L. Chen et al. (Eds.): DASFAA 2009 Workshops, LNCS 5667, pp. 6–20, 2009.

labeling scheme' we mean a labeling scheme where we can extract all the ancestor's labels from a node label.

Path labeling schemes have generally better update features [19,17]. Moreover, a path labeling scheme such as OrdPath can be updated without any relabeling [17]. A path labeling scheme has variable length labels, however, this problem can be solved using simple label encoding [7]. Despite these features we can not find many efficient query processing algorithms using path labeling schemes. Existing algorithms using structural joins [22,1,6] or holistic joins [3,5,4] work only with an element labeling scheme.

Holistic algorithms are designed for specific types of XPath queries; so called twig query patterns. However, the holistic algorithm can be understood as an operator of XML algebra [15] and therefore it can be utilized in the case of more complex XPath queries. In [16], we can find a comparison of approaches to a twig query processing. Holistic algorithms were considered as the most robust solution not requiring any complicated query optimizations. Moreover, holistic approaches provide a theoretical model for optimal processing of some query classes and, their main memory requirements are minimal in this case.

The TJFast holistic algorithm [13] applies the Extended Dewey path labeling scheme, but the update features are constrained by a finite state transducer used with Extended Dewey. Moreover, TJFast must extract the labeled path from every label and compare it with the regular expression.

We improve the TJFast holistic algorithm using the DataGuide summary tree in order to decrease the unnecessary computing cost and I/O cost significantly. Our work shows, that the query algorithm using a path labeling scheme can outperform existing state-of-the-art algorithms using an element labeling scheme [3,5,4]. Due to the fact that our algorithm can be used with a popular Dewey order labeling scheme we call it TJDewey. However, the TJDewey algorithm is not dependent on a specific path labeling scheme, it can be used with *any* path labeling scheme.

This paper is organized as follows. In Section 2, we depict a model of an XML document. Section 3 focuses on a brief description of holistic approaches. Due to the fact that the TJFast method is in the scope of our paper, we describe this approach in more detail. In Section 4, we introduce the improvement of TJFast. Section 5 provides comprehensive experimental results of different holistic approaches. In the last section, we summarize the paper content and outline possibilities of our future work.

2 Model

An XML document can be modeled as a rooted, ordered, labeled tree, where every node of the tree corresponds to an element or an attribute of the document and edges connect elements or elements and attributes having a parent-child relationship. We call such representation of an XML document an *XML tree*. We see an example of the XML tree in Figure 3(b). We use the term 'node' in the meaning of a node of an XML tree which represents an element or an attribute.

For each node of an XML tree, we shell define a *labeled path* as a sequence $tag_0/tag_1/\ldots/tag_n$ of node tags lying on a path from the root to the node n. A labeled path provides additional information about a node that can be utilized to speed