

ADVANCES ON
INFORMATION
PROCESSING AND
MANAGEMENT

CONFERENCE ORGANIZERS INSTITUTES

The International Conference on Integrated Information is supported by the following Institutes:

Emerald Group Publishing Limited
Technological educational Institute of Athens, Greece
University of Peloponnese, Greece
National And Kapodistrian University of Athens, Greece
Mednet Hellas, The Greek Medical Network
2nd AMICUS Workshop

To learn more about I-DAS, including the Book Series, please visit the webpage
<http://www.i-das.org/>

INTEGRATED INFORMATION

International Conference on Integrated Information

Kos, Greece September, 29 – October, 3 2011

EDITORS

Georgios A. Giannakopoulos
Technological Educational Institute of Athens, Greece

Damianos P. Sakas
University of Peloponnese, Greece

All papers have been peer-reviewed



Piraeus, Greece, 2011

Editors

Georgios A. Giannakopoulos

Technological Educational Institute of Athens
Faculty of Management and Economics
Department of Library Science and Information Systems
Address: Aghiou Spyridonos Street, 12210, Egaleo
E-mail: gian@teiath.gr

Damianos P. Sakas

University of Peloponnese
Faculty of Science and Technology
Department of Computer Science and Technology
Address: End of Karaiskaki St., 22100, Tripolis, Greece
E-mail: D.Sakas@uop.gr

The copyrights will be owned by the authors under the Creative Commons Attribution-Non Commercial license (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted use, distribution, and reproduction in any non commercial medium, provided the original work is properly cited.

ISSN:

Printed in the Greece, EU

CONTENTS

PREFACE: Proceedings of the International Conference on Integrated Information (IC-INFO 2011)	1
Georgios A. Giannakopoulos, Damianos P. Sakas	
Conference Details	3
Keynote Speaker	5
SYMPOSIUM ON INFORMATION AND KNOWLEDGE MANAGEMENT	6
Prof. Christos Skourlas	
Towards the Preservation and Availability of Historical Books and Manuscripts: A Case Study	8
Eleni Galiotou	
An Extensive Experimental Study on the Cluster-based Reference set Reduction for Speeding-up the k-nn Classifier	12
Stefanos Ougiaroglou, Georgios Evangelidis and Dimitris A. Dervos	
Exploiting the Search Culture Modulated by the Documentation Retrieval Applications	16
Nikitas N. Karanikolas and Christos Skourlas	
Information and Knowledge Organization: The Case of the TEI of Athens	22
Anastasios Tsolakidis, Manolis Chalaris and Ioannis Chalaris	
Providing Access to Students with Disabilities and Learning Difficulties in Higher Education through a Secure Wireless framework	26
Catherine Marinagi and Christos Skourlas	
Improving Query Efficiency in High Dimensional Point Indexes	30
Evangelos Outsios and Georgios Evangelidis	
Text Segmentation Using Named Entity Recognition and co-Reference Resolution in Greek Texts	34
Pavlina Fragkou	
KINISIS, a Graphical XQuery Language	42
Euclid Keramopoulos, Achilleas Pliakas, Konstantinos Tsekos and Ignatios Deligiannis	
Dimensionality Curse, Concentration Phenomenon and the KDB-tree	46
Nikolaos Kouiroukidis and Georgios Evangelidi	

Applying Balanced Scorecard Strategic Management in Higher Education	50
Manolis Chalaris, Anastasios Tsolakidis and Ioannis Chalaris	
A Web Portal Model for NGOs' Knowledge Management	54
Zuhal Tanrikulu	
The Digital Archives System and Application Optimized for the Tradition Knowledge Archives	58
Jeon Hong. Chan, In Deok. Hwang, Jae Hak. Park, Hyeok. Sim, U won. Gwon and Soon Cheol. Park	
A Semi-automatic Emerging Technology Trend Classifier Using SCOPUS and PATSTAT	62
Seonho Kim, Woondong Yeo, Byong-Youl Coh, Waqas Rasheed, Jaewoo Kang	
Presenting a Framework for Knowledge Management within a Web Enabled Living Lab	66
Lizette de Jager and Albertus AK Buitendag and Potjie (JS) van der Walt	
4TH SYMPOSIUM ON BUSINESS AND MANAGEMENT AND DYNAMIC SIMULATION MODELS SUPPORTING MANAGEMENT STRATEGIES	71
Dr. Damianos Sakas	
New Political Communication Practices: No Budget Events Management. The New Challenge	73
Evangelia N. Markaki, Damianios P. Sakas and Theodore Chadjipantelis	
Free Software – Open Source Software. A Powerful Tool for Developing Creativity in the Hands of the Student	78
Nasiopoulos K. Dimitrios, Damianos P. Sakas, Konstantinos Masselos	
Open Source Web Applications. How it Spread Through the Internet and their Contribution to Education.	82
Nasiopoulos K. Dimitrios, Damianos P. Sakas, Konstantinos Masselos	
Culture in Modern Times in the Frame of Luhmann's System Theory	85
Anastasia J. Chournazidis	
Managing Scientific Journals: A Cultural Viewpoint	87
Marina C. Terzi, Damianos P. Sakas, and Ioannis Seimenis	
A Conceptual Framework for Analyzing Knowledge-based Entrepreneurship	92
Nikos S. Kanellos	

SESSION ON INFORMATION HISTORY: PERSPECTIVES, METHODS AND CURRENT TOPICS	96
Prof. Laszlo Karvalics	
Emerging Research Fields in Information History	98
Laszlo Z. Karvalics	
Information Management through Elementary Data Clusters: New Observations on Pridianum-Type Roman Statistical Documents	102
Gergő Gellérfi	
Information and Secrecy on the Silk Road. Methods of Encryption of Legal Documents in Inner Asia (3th-4th century)	106
Szabolcs Felföldi	
The Role of Information and Disinformation in the Establishment of the Mongolian Empire: A Re-examination of the 13th century Mongolian History from the Viewpoint of Information History	110
Márton Gergő Vér	
Early Warning Systems and the Hospitallers in the Eastern Mediterranean	114
Zsolt Hunyadi	
Information Management as Establishment Dutch Navigational Knowledge on Japan, 1608-1641	118
Gabor Szommer	
Files Everywhere - Register and Training of Men for Military and Civil Purpose in Prussia in the early 18th century	123
Marton Holczer	
SYMPOSIUM ON INTEGRATED INFORMATION: THEORY, POLICIES, TOOLS	126
Prof. Georgios Giannakopoulos	
Approaching Information as an Integrated Field: Educating Information Professionals	128
Georgios Giannakopoulos, Daphne Kyriaki Manesi and Stryidon Zervos	
Special Libraries as Knowledge Management Centers	132
Eva Semertzaki	
Digital Libraries' Developers and their Suitability: A Case Study	136
Maria Monopoli	

A Preliminary Study for the Creation of a Greek Citation index in the Humanities and the Social Sciences (GCI – H&SS)	140
Daphne Kyriaki-Manessi and Evi Sachini	
Archiving as an Information Science. Evidence from a Survey Carried out on a Sample of Greek Students	144
Georgios Giannakopoulos and Ioannis Koumantakis	
Transition Process of E-records Management and Archiving System in Universities: Ankara University	147
S. Özlem Bayram and Fahrettin Ozdemirci	
Government Information: Access and Greece's Efforts for Access	150
Aikaterini Yiannoukakou	
School Archives and their Potentials in Teaching: Aspects of Greek Reality	156
Sonia Geladaki and Panagiota Papadimitriou	
Research on School Libraries in Greece and Suggestions on its Further Development	160
Georgios D. Bikos	
Building Digital Collections for Archeological Sites: Metadata Requirements and CIDOC CRM Extension	164
Georgios S. Gkrous and Mara Nikolaidou	
Museological Claims to Autonomous Knowledge: Rethinking the Conceptual Mode of Display and its Claims to Knowledge	169
Assimina Kaniari and Georgios Giannakopoulos	
Use of Library Loan Records for Book Recommendation	172
Keita Tsuji, Erika Kuroo, Sho Sato, Ui Ikeuchi, Atsushi Ikeuchi, Fuyuki Yoshikane and Hiroshi Itsumura	
Developing a National Database on Librarianship and Information Science. The Case of E-VIVA, the Hellenic Fulltext Database	176
Filippos Ch. Tsimpoglou, Vasiliki V. Koukounidou and Eleni K. Sakka	
Integrated Access to Cultural Heritage Information Pieces in Iran Astan-Quds Razavi's Organization of Libraries, Museums and Documents Center: A Theory of Unionization Disparate Information Assets over Imam Reza's Zarih	181
Ms. Mitra Zarei and Ms. Maliheh Farrokhnia	
Attitudes of University Librarians and Information Scientists towards the Draft Code of	185

Library Ethics to Present a Model for Final Library Ethical Codes Mahsoomeh Latifi, Fatemeh Zandian and Hasan Siamian	
SESSION ON OPEN ACCESS REPOSITORIES: SELF-ARCHIVING, METADATA, CONTENT POLICIES, USAGE Dr. Alexandros Koulouris	188
Geographical Collections in Greek Academic Libraries: Current Situation and Perspectives Ifigenia Vardakosta and Sarantos Kapidakis	189
Information Seeking Behavior: Factors that Affect the Behavior of Greek Astronomers Hara Brindesi and Sarantos Kapidakis	194
Aggregating Metadata for Europeana: The Greek Paradigm Alexandros Koulouris, Vangelis Banos and Emmanouel Garoufallou	198
Integrating a Repository with Research Output and Publications: The Case of the National Technical University of Athens Dionysis Kokkinos	202
Implementation of Workflows as Finite State Machines in a National Doctoral Dissertations Archive Nikos Houssos, Dimitris Zavaliadis, Kostas Stamatis and Panagiotis Stathopoulos	205
Practices of “Local” Repositories of Legally Protected Immovable Monuments. A Global Scheme for ‘Designation – Significance’ Information Michail Agathos and Sarantos Kapidakis	209
Integration of Metadata in BWMETA-2.0.0 Format Katarzyna Zamlynska, Jakub Jurkiewicz and Lukasz Bolikowski	213
SESSION ON EVIDENCE-BASED INFORMATION IN CLINICAL PRACTICE Dr. Evangelia Lappa	216
Applicability of Data Mining Algorithms on Clinical Datasets Wilfred, Bonney	218
Changing Roles of Health Librarians with Open Access Repositories Christine Urquhar and Assimina Vlachaki	221
From Medical Records to Health Knowledge Management Systems: The Coding to Health Sector Evangelia C. Lappa and Georgios A. Giannakopoulos	225

The Survey of Skill, Attitude and Use of Computer and Internet among Faculty Members	229
Hasan Siamian, Azita Bala Ghafari, Kobra Aligolbandi, Mohammad Vahedi and Gholam Ali Golafshani Jooybari	
Trends in Scholarly Communication among Biomedical Scientists in Greece	232
Assimina Vlachaki and Christine Urquhart	
SESSION ON ELECTRONIC PUBLISHING: A DEVELOPING LANDSCAPE	236
Dr. Dimitris Kouis	
E-Journal and Open Access Journal Publishing in the Humanities: Preliminary Results from a Survey among Byzantine Studies Scholars	238
Victoria Tsoukala and Evi Sachini	
Preliminary Results on a Printed VS Electronic Text Books Assessment Through Questionnaire	242
Dimitrios A. Kouis and Kanella Pouli	
An Interpretation of Aristotelian Logic According to George Boole	246
Markos N. Dendrinis	
SESSION ON INFORMATION CONTENT PRESERVATION AS OUTCOME OF CONSERVATION OF CULTURAL HERITAGE: ETHICS, METHODOLOGY AND TOOLS	251
Prof. George Panagiaris and Dr. Spiros Zervos	
Intrinsic Data Obfuscation as the Result of Book and Paper Conservation Interventions	254
Spiros Zervos, Alexandros Koulouris and Georgios Giannakopoulos	
Mass Deacidification: Preserving More than Written Information	258
Michael Ramin, Evelyn Eisenhauer and Markus Reist	
Information Literacy of Library Users: A Case Study of Mazandaran Public Library Users, Iran	260
Hussein Mahdizadeh and Hasan Siamian	
The Narratives of Paper in The Archives of the New Independent Greek State (Mid 19th c.)	264
Ourania Kanakari and Maria Giannikou	
From Macro to Micro and from Micro to Nano: The Evolution of the Information Content Preservation of Biological Wet Specimen Collections	268
Nikolaos Maniatis and Georgios Panagiaris	

Digital images: A valuable scholar's tool or misleading material?	272
Patricia Engel	
Attitudes of University Librarians and Information Scientists Towards the Draft Code of Library Ethics to Present a Model for Final Library Ethical Codes	277
Mahsoomeh Latifi, Fatemeh Zandianand and Hasan Siamian	
Investigation of the Degradation Mechanisms of Organic Materials: From Accelerated Ageing to Chemometric Studies	280
Ekaterini Malea, Effie Papageorgiou and Georgios Panagiaris	
SESSION ON DIVERGENCE AND CONVERGENCE: INFORMATION WORK IN DIGITAL CULTURAL MEMORY INSTITUTIONS	285
Dr. Susan Myburgh	
Extending Convergence and Divergence in Cultural Memory Institutions: The Old Slave Lodge in the New South Africa	287
Archie L Dick	
The Transfer of Knowledge from Large Organizations to Small: Experiences from a Research Project on Digitization in Wales	289
Clare Wood-Fisher, Richard Gough, Sarah Higgins, Menna Morgan, Amy Staniforth and Lucy Tedd	
The Usage of Reference Management Software (Rms) in an Academic Environment : A Survey at Tallinn University	293
Enrico Francese	
Varialog : How to Locate Words in a French Renaissance Virtual Library	297
Marie-Hélène Lay	
The Urge to Merge: A Theoretical Approach	301
Susan Myburgh	
SYMPOSIUM ON ADVANCES INFORMATION FOR STRATEGIC MANAGEMENT	304
Professor Nikolaos Konstantopoulos	
Empowerment in the Tax Office of Greece	306
Antonios E. Giokas and Nikolaos P. Antonakas	
Building Absorptive Capacity Through Internal Corporate Venturing	310
Ioannis M. Sotiriou and Alexandros I. Alexandrakis	

The Monitoring Information System (M.I.S.) - An information and Management System for Projects Co-financed Under the National Strategic Reference Framework (NSRF) and the Community support framework (CSF)	314
Catherina G. Siampou, Eleni G. Fassou and Athanassios P. Panagiotopoulos	
Corruption in Tax Administration: The Entrepreneurs View Point	318
Nikolaos P. Antonakas, Antonios E. Giokas and Nikolaos Konstantopoulos	
Conflicts between the IT Manager and the Software House after the Strategic Choice of Outsourcing of the Information Processes in Maritime Companies.	322
Anthi Z. Vaxevanou, Nikolaos Konstantopoulos, Damianos P. Sakas	
Contemporary Forms of Ordering Between the Supply Department and Ship Chandler Companies in the Shipping Industry	325
Anthi Z. Vaxevanou, Nikolaos Konstantopoulos, Damianos P. Sakas	
Strategies Implemented and Sources Used for the Acquisition of Information on Foreign Markets	329
Myropi Garri, Nikolaos Konstantopoulos and Michail G. Bekiaris	
The Effect of High Performance Working Systems on Informative Technology in Enterprises after Organisation Changes such as Mergers & Acquisitions	333
Nikolaos Konstantopoulos and Yiannis Triantafyllopoulos	
Personnel's Absorptive Capacity as a Guiding Concept for Effective Performance in Informative Technology	337
Nikolaos Konstantopoulos and Yiannis Triantafyllopoulos	
SESSION ON CONTEMPORARY ISSUES IN MANAGEMENT: ORGANISATIONAL BEHAVIOUR, INFORMATION TECHNOLOG, EDUCATION & HOSPITAL LEADERSHIP	341
Dr. Panagiotis Trivellas	
Investigating the Importance of Sustainable Development for Hotel SMES	343
Panagiotis Reklitis and Anestis Fotiadis	
Strategic Alignment of ERP, CRM and E-business: A Value Creation	347
Catherine C. Marinagi and Christos K. Akrivos	
The Impact of Occupational Stress on Performance in Health Care	351
Panagiotis Trivellas Panagiotis Reklitis and Charalambos Platis	

The Impact of Emotional Intelligence on Job Outcomes and Turnover Intention in Health Care	356
Panagiotis Trivellas Vassilis Gerogiannis and Sofia Svarna	
SYMPOSIUM ON BUSINESS MANAGEMENT AND COMMUNICATION STRATEGIES SUPPORTING DECISION MAKING PROCESS IN TOURISM SECTOR	360
Dr. Panagiota Dionysopoulou	
The Human Factor as a Mediator to the Total Quality in the Tourism Companies. The impact of Employees' Motivation to Quality Improvements	362
Christos K. Akrivos and Panagiotis Reklitis	
Tourist Destination Marketing and Management Using Advanced ICTS Technologies	365
Anastasia Argyropoulou, Panagiota Dionyssopoulou, Georgios Miaoulis	
G.N.T.O. (Greek National Tourism Organization) Communication Strategy in Advertising Campaigns 1991-2006	370
George Stafylakis and Panagiota Dionyssopoulou	
GENERAL PAPERS	375
The role of Environmental Education within the Framework of the Environmental Policy of a Regional Municipality	376
Vassiliki Delitheou and Dimitra Thanasia	
Issues of Social Cohesion: A case study from the Greek Urban Scenery	380
Evgenia Tousi	
Merging Activity and Employee Performance: The Greek Banking System	384
Panagiotis Liargovas and Spyridon Repousis	
Sustainable Development and Corporate Social Responsibility in Higher Education: Some Evidence from Greece	387
Anastasios Sepetis and Fotios Rizos	
Exploring the Effects of Organizational Culture on Collaborative vs. Competitive Knowledge Sharing Behaviors	395
Hanan Abdulla Mohammed Al Mehairi and Norhayati Zakaria	

Preface: Proceedings of the International Conference on Integrated Information (IC-ININFO 2011)

GEORGIOS A. GIANNAKOPOULOS

Department of Library Science and Information Systems, Technological Educational Institute of Athens, Aghiou Spyridonos, Egaleo, 12210, Greece

DAMIANOS P. SAKAS

Department of Computer and Technology Science, University of Peloponnese, Praxitelous 89-91, Piraeus, 18532, Greece

Aims and Scope of the Conference

The International Conference on Integrated Information 2011 took place in Kos Island, Greece, between September, 29 and October, 3, 2011. IC-ININFO is an international interdisciplinary conference covering research and development in the field of information management and integration.

The conference aims at creating a forum for further discussion for an Integrated Information Field incorporating a series of issues and/or related organizations that manage information in their everyday operations. Therefore, the call for papers is addressed to scholars and/ or professionals of the fields of Library and Archives Science (including digital libraries and electronic archives), Museum and Gallery Studies, Information Science, Documentation, Information Management, Records Management, Knowledge Management, Data management and Copyright experts the latter with an emphasis on Electronic Publications. Furthermore, papers focusing on issues of Cultural Heritage Management and Conservation Management are also be welcomed along with papers regarding the Management of Nonprofit Organizations such as libraries, archives and museums.

One of the primary objectives of the IC-ININFO will be the investigation of information-based managerial change in organizations. Driven by the fast-paced advances in the Information field, this change is characterized in terms of its impact on organizations that manage information in their everyday operations.

Grouping emerging technologies in the Information field together in a close examination of practices, problems and trends, IC-ININFO and its emphases on integration and management will present the state of the art in the field. Addressed jointly to the academic and practitioner, it will provide a forum for a number of perspectives based on either theoretical analyses or empirical case studies that will foster dialogue and exchange of ideas.

Topics of general Interest

Library Science, Archives Science, Museum and Gallery Studies, Information Science, Documentation, Digital Libraries, Electronic Archives, Information Management, Records / Document Management, Knowledge Management, Data Management, Copyright, Electronic Publications, Cultural Heritage Management, Conservation Management, Management of Nonprofit Organizations, History of Information, History of Collections, Health Information

Symposia

The Conference offered a number of sessions under its patronage, providing a concise overview of the most current issues and hands-on experience in information-related fields.

- Symposium on Integrated information: Theory, Policies, Tools
- 4th Symposium on Business and Management and Dynamic Simulation Models supporting management strategies

- Session on Open Access Repositories: Self-archiving, Metadata, Content policies, Usage
- Session on Evidence-Based Information in Clinical Practice
- Session on Business Management and Communication Strategies supporting Decision Making Process in Tourism Sector
- Session on Electronic Publishing: A Developing Landscape
- Session on Information and Knowledge Management
- Session on Information Content Preservation as Outcome of Conservation of Cultural Heritage: Ethics, Methodology and Tools
- Session on Advances Information for Strategic Management
- Session on Information History: Perspectives, Methods and Current Topics
- Session on Divergence and Convergence: Information Work in Digital Cultural Memory Institutions
- Session on Contemporary issues in Management: Organisational Behaviour, Information Technology, Education & Hospital leadership.

The wide range of aspects that the sessions covered, highlighted future trends in the Information Science.

Paper Peer Review

More than 300 papers had been submitted for consideration in IC-ININFO 2011. From them, 91 were selected for presentation, after peer review in a double blind review process. The accepted papers were presented at IC-ININFO 2011.

Thanks

We would like to thank all members that participated in any way in the IC-ININFO 2011 Conference and especially:

- The famous publishing house Emerald for its communication sponsorship.
- The co-organizing Universities and Institutes for their support and development of a high-quality Conference scientific level and profile.
- The members of the Scientific Committee that honored the Conference with their presence and provided a significant contribution to the review of papers as well as for their indications for the improvement of the Conference.
- All members of the Organizing Committee for their help, support and spirit participation before, during and after the Conference.
- The Session Organizers for their willing to organize sessions of high importance and for their editorial work, contributing in the development of valued services to the Conference.
- PhDC Marina Terzi for her excellent editorial work, contributing in the production of the Conference proceedings.

CONFERENCE DETAILS

Chairs

Georgios A. Giannakopoulos, Technological Educational Institute of Athens, Greece
Damianos P. Sakas, University of Peloponnese, Greece

Co-Chairs

Daphne Kyriaki – Manesi, Technological Educational Institute of Athens, Greece
Dimitrios Vlachos, University of Peloponnese, Greece

Scientific Committee

Amanda Spink, Queensland University of Technology
Andreas Bagias, European Court
Andreas Rauber, Vienna University of Technology
Astrid van Wesenbeeck, SPARC Europe
Christine Urquhart, Aberystwyth University
Christos Schizas, University of Cyprus
Christos Skourlas, Technological Educational Institute of Athens
Claire Farago, University of Colorado at Boulder
Claus-Peter Klas, FernUniversität in Hagen
Costas Vassilakis, University of Peloponnese,
Dimitris Dervos, Technological Educational Institute of Thessaloniki
Eelco Ferwerda, OAPEN
Elena Garcia Barriocanal, University of Alcalá
Emmanuel Garoufallou, Technological Educational Institute of Thessaloniki
Filippos Tsimpoglou, University of Cyprus
Fillia Makedon, University of Texas at Arlington
George Korres, University of Newcastle
Georgios Evangelidis, University of Macedonia
Georgios Panagiaris, Technological Educational Institute of Athens
Johan Oomen, Netherlands Institute for Sound and Vision
José Aldana, University of Malaga
Konstantinos Masselos, University of Peloponnese
Luciana Duranti, The University of British Columbia
Markos N. Dendrinis, Technological Institute of Athens
Milena Dobрева, University of Strathclyde
Prodromos Tsiavos, London School of Economics and Political Science
Sándor Darányi, University of Borås
Sarantos Kapidakis, Ionian University
Sirje Virkus, Tallinn University
Spiros Zervos, Technological Educational Institute of Athens
Susan Myburgh, University of South Australia
Theodoros Pitsios, University of Athens, Faculty of Medicine

Organizing Committee

Alexandros Koulouris (Chair), Technological Educational Institute of Athens
Christos Christopoulos, SCEV Scientific Events Ltd
Marina Terzi, University of the Aegean, Greece
Evangelia Markaki, Aristotle University of Thessaloniki

Assimina Kaniari, Athens School of Fine Arts
Evangelia Lappa, General Hospital Attikis K.A.T.
Dimitris Kouis, Greek Ministry of Education, Lifelong Learning and Religious Affairs
Dionysis Kokkinos, National Technical University of Athens

KEYNOTE SPEAKER



Professor Amanda Spink

Professor Amanda Spink has published over 340 scholarly journal articles, refereed conference papers and book chapters, and 6 books. Many of her journal articles are published in the Journal of the American Society for Information Science and Technology, Information Processing and Management, and the Journal of Documentation. She is Editor of the Emerald journal Aslib Proceedings. Amanda's research has been published at many conferences including ASIST, IEEE ITCC, CAIS, Internet Computing, ACM SIGIR, and ISIC Conferences. Her recent books include Information Behavior: An Evolutionary Instinct and Web Search: Multidisciplinary Perspectives, both published by Springer. Amanda's research focuses on theoretical and empirical studies of information behavior, including the evolutionary and developmental foundations. The National Science Foundation, the American Library Association, Andrew R. Mellon Foundation, Amazon.com, Vivisimo. Com, Infospace.com, NEC, IBM, Excite.com, AlltheWeb.com, AltaVista.com, FAST, and Lockheed Martin have sponsored her research. In 2008 Professor Spink had the second highest H-index citation score in her field from 1998 to 2008 [Norris, M. (2008)]. Ranking Fellow Scholars and their H-Index: Preliminary Survey Results. Loughborough University, Dept of Information Science Report].

Improving Query Efficiency in High Dimensional Point Indexes

Evangelos Outsios and Georgios Evangelidis

University of Macedonia, Department of Applied Informatics, 54006, Thessaloniki, Greece
{outsios, gevan} (at) uom.gr

Abstract: *In this paper, we focus on the leaf level nodes of tree-like k-dimensional indexes that store the data entries, since those nodes represent the majority of the nodes in the index. We propose a generic node splitting approach that defers splitting when possible and instead favors merging of a full node with an appropriate sibling and then re-splitting of the resulting node. Our experiments with the hB-tree, show that the proposed splitting approach achieves high average node storage utilization regardless of data distribution, data insertion patterns and dimensionality.*

Keywords: *K-dimensional point indexing, Optimizing data node storage utilization, Range query performance*

I. INTRODUCTION

Lately, with the increased interest in Data Mining, indexing of k-dimensional vectors has become essential when dealing with kNN classification. Brute force application of kNN classification on large databases involves as many computations of distances as the size of the database, since one has to find the k closest points to the query point. Data reduction and/or data dimensionality reduction techniques are used to reduce the computational cost, but they usually decrease the accuracy of the kNN classifier. Alternatively, indexing can be used to reduce the linear cost of searching to logarithmic. Unfortunately, all high-dimensional indexes suffer from the “dimensionality curse” problem. It has been shown that above 8 dimensions, most indexes perform no better than the exhaustive sequential search of the whole database when answering kNN queries (Berchtold et al., 1998).

For very large high-dimensional datasets, the most sensible approach to kNN classification is a combination of a data dimensionality reduction technique, to reduce the dimensions down to 8 to 16, and then, the use of a high dimensional point index. That is why the quest for efficient indexes in medium to low dimensions has regained the interest of the research community. Efficient indexes should not be affected by the cardinality of the dataset, the data distribution, the dimensionality, and the insertion patterns. Since the kNN classifier is a model-free classifier, new insertions in the dataset should dynamically update the model, i.e., the index, without affecting its performance.

Indexes with guarantees in node storage utilization, obviously, lead to better query performance, since fewer nodes (disk pages) are visited to answer a query. kNN queries are a specialization of range queries and require

visiting of multiple leaf or data level nodes of the index, where rids of the points or the points themselves are stored.

In this paper, we deal with tree-like k-dimensional indexes that partition the space in non-overlapping subspaces, like the KDB-tree (Robinson, 1981) or the hB-tree (Lomet and Salzberg, 1990; Evangelidis et al., 1997). The hB-tree and the hB-pi* tree (Zhou and Salzberg, 2008), a variation that also indexes empty space, have been recently shown to outperform the R*-tree (Beckmann et al., 1990), the most well known spatial index. We focus on their leaf or data level index nodes since those nodes represent the majority of the index nodes. We propose a generic node splitting approach that delays data node splitting when possible and instead favors redistribution of the contents of a full node with an appropriate sibling. Our experiments with the hB-tree, show that the proposed splitting approach achieves high node storage utilization and good range query performance.

In Section II, we present related work in improving data node storage utilization and provide a short description of the KDB-tree and the hB-tree. We also present a policy for selecting the splitting attribute in high dimensional indexes. We propose a new data node splitting method in Section III, and we present experimental results in Section IV. Finally, we conclude the paper in Section V.

II. RELATED WORK

In this section, we review the approaches that have been proposed in the literature for improving storage utilization. First, we discuss the 1-dimensional case with the B+tree, and then, we briefly describe the structure of the hB-tree and the KDB-tree. Finally, we give some insight on how splitting attribute selection policies can improve storage utilization and range query performance, when splitting data nodes.

A. Data node storage utilization

For the B+tree, there are many ways to increase node storage utilization (Comer, 1979). For example, Knuth (1973), proposes to delay splitting by locally redistributing the contents of nodes until two sibling nodes become full. Then the two full nodes are split into three nodes with a node storage utilization of at least 66%, an improvement over the 50% storage utilization of the B+tree. Although the average node storage utilization remains unaffected and about 69% ($\frac{1}{2} \ln 2$) for uniform data distributions (Yao, 1978), this approach achieves better storage utilization for non-uniform data distributions.

In addition, even for uniform data distributions, index performance is affected by the way data points are inserted in the index. For random (uniform) insertion patterns there is no difference on the way nodes are split. As long as nodes are split in a 1:1 ratio, average storage utilization is close to 69%. But under different patterns of insertion, for example, block insertions, where incoming points are inserted to a particular node until that node splits, average storage utilization can degrade considerably.

The picture is quite different when indexing in high dimensions. Almost all of the proposed k-dimensional indexes do not provide such guarantees. Only the hB-tree, that splits its nodes in a 1:2 ratio in the worst case, achieves a comparable to the B+tree average node storage utilization (about 67%) and worst node storage utilization of 33%. But under certain patterns of record insertions, even the hB-tree can have average node utilization close to 50%. In the k-dimensional paradigm, it is not always possible to merge and re-split nodes as in the 1-dimensional case of the B+tree, because the notions of the “next” and “previous” sibling nodes cannot be defined. Redistribution of entries among nodes is much more complicated, and, depending on the index at hand, involves complicated updates on the corresponding index terms of the participating nodes.

B. KDB-tree and hB-tree

In both the KDB-tree and the hB-tree, data nodes, i.e., leaf level index nodes, contain the k-dimensional points or data terms for those points (in the case of secondary indexes). In a way analogous to the B+tree, when a data node becomes overfull because of insertions of new points, it has to be split. After the split we end up with two data nodes, the initial one occupying the same disk page and a new one occupying a new disk page. This process is repeated continuously, every time a data node becomes overfull.

The KDB-tree splits data nodes always using a single attribute, thus all data nodes are hyper-rectangles (or bricks). Also, internal nodes, i.e., index nodes above the leaf level, are split either at the root of their internal kd-tree, thus by a hyperplane, or by using some other splitting attribute to achieve balanced splits at the cost of downward propagation of splits. The KDB-tree does not have any guarantees on node storage utilization.

The hB-tree is an improvement of the KDB-tree, since it guarantees an average node storage utilization of 67% by splitting nodes at a 1:2 ratio in the worst case (compare this with B+tree’s 1:1 ratio). This can be achieved both in the internal nodes that contain index terms in the form of kd-trees and in the data nodes that contain data entries. In Lomet and Salzberg (1990), it is shown that a 1:2 split ratio is always possible. In internal nodes, an appropriate kdsbtree is extracted from the overfull node. In data nodes, it may be necessary to use more than one attributes to achieve such a split. The overfull node and the newly extracted node can be hyper-rectangles from which smaller hyper

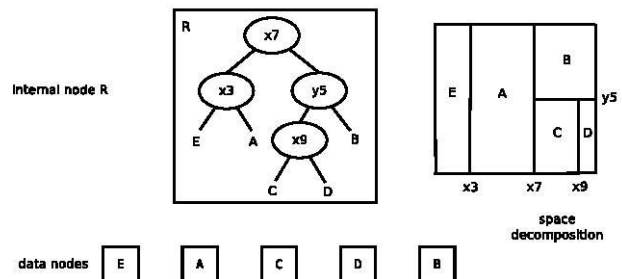


Figure 1: An example hB-tree

rectangles have been extracted, thus the name holey-Brick-tree (hB-tree). -

In Figure 1, an hB-tree with two levels is shown. It contains 5 data nodes and an internal node R, that is the root of the hB-tree. R contains the index terms for its 5 children in the form of a kd-tree. Let’s assume that the last split that happened was the one that extracted node E from node A. Before the split, kd-tree node x7 in R had a left pointer to data node A. The index term consisting of kd-tree node x3 (namely, the attribute and attribute value that were used to split A and extract E) was merged in the kd-tree of R to describe the new space decomposition.

C. Splitting attribute selection policy

When splitting overfull data nodes the goal is (a) to minimize the cost of future range queries, and, (b) to maximize average node storage utilization. The second goal, although it creates smaller trees, it may conflict with the first goal. This is because, good node storage utilization can lead to poor k-dimensional space partitioning.

For good space partitioning, the obvious approach is to split the space of the data node in half along the longest edge (attribute) and to ignore the distribution of the points in the node. The resulting data nodes will index the same amount of space and will have regular shapes, i.e., edges of similar lengths. Thus, they will have the same probability of receiving new insertions of points or of being visited by subsequent range queries. The drawback of this approach is that we may end up having nodes with low or zero storage utilization. Alternatively, one may choose to achieve the best possible node storage utilization by always trying to achieve 1:1 point splits, at the cost of bad space partitioning.

In Outsios and Evangelidis (2010), we experimented with various splitting attribute selection policies for data nodes. In this paper we choose the policy that uses the best attribute for even point split and best possible space split. This works as follows. Choose the attribute that achieves the most even point split. In case of ties, choose the attribute that splits along the longest edge. By splitting along the longest edge, we favor hyper-rectangles that are as close as possible to hyper-cubes. The goal is to minimize the cost of range queries by avoiding peculiar shaped subspaces.

III. NEW SPLITTING APPROACH

We focus our attention to the leaf level of the index, since the data nodes are the majority of the nodes in the tree index.

When splitting data nodes we should aim at:

1. Splitting the data node as evenly as possible both in terms of points (to improve node storage utilization) and space (to improve range query performance). We achieve this by using the best attribute for even point split, and, at the same time, the best possible space split.
2. Posting the most compact index term possible to minimize the number of the internal index nodes. We achieve this by always performing hyperplane splits. Thus, we minimize the size of the index terms, and the resulting data nodes are always hyper-bricks.

To further improve the performance under non-uniform data insertion patterns, we propose a new way for dealing with overfull nodes. We first define the terms paired and single data nodes. To illustrate the term paired, we examine Figure 1. Data nodes E and A are paired since they are pointed by the same kd-tree node in the kd-tree of their parent R. Data nodes C and D are also paired, whereas, data node B is considered to be single.

The idea is to exploit the structure of the kd-tree in the internal index nodes right above the data nodes, in order to identify data nodes that can re-distribute their contents. Following such an approach, leads to delayed splits of overfull nodes until their paired node becomes overfull, too.

IV. EXPERIMENTAL EVALUATION

We tested our splitting approach against the standard splitting algorithm of the hB-tree. The tested variations were the following:

- m1 Original hB-tree data node splitting algorithm: do not use any redistribution scheme. When a node becomes full, split it.
- m2 Redistribute among paired nodes and eventually split: Paired nodes delay splitting by redistributing their contents with their paired sibling. Only when both nodes in a pair become full, the one that overflows, splits. Single nodes split when full.

Parameter	Values
attribute value range	[0, 1]
k=dimensionality	2 – 15
database size	100K points
DNS=data node sizes	10 – 100 points
INS=internal node sizes	5 – 50 kd-tree nodes
insertion patterns	uniform and block
range query space selectivity	0.01%

Table 1: Experiment parameters and values

Table 1 lists the parameters of the experiments and the values we used.

We used relatively small node sizes in order to build hB-trees with many levels and stress our algorithms. Notice that the data node size is affected by the dimensionality of the points, i.e., 10 2-dim points occupy 1/10th of the space occupied by 10 20-dim points.

Also, we assumed that there are no deletions of points. The index only grows in time. To achieve the desired range query windows, we generated hypercube queries that covered 0.01% of the k-dimensional space. Thus, for 100K uniformly distributed points, we expect the query window to contain 10 points.

In Table 2, we compare the splitting methods m1 and m2 on average data node storage utilization and range query efficiency (in terms of average number of visited pages to answer 100 random queries with 0.01% selectivity).

We use 100K uniformly distributed points with uniform insertion pattern and we vary dimensionality. Using small node sizes we build trees with 7 levels. We observe that m1 and m2 have comparable node storage utilization across dimensions, but, as expected, m2 builds slightly smaller trees, i.e., with fewer data nodes. Thus, m2 performs slightly better in terms of average data node storage utilization and average number of accessed nodes per range query.

Next, we focus on node storage utilization when using a block insertion pattern, i.e., when a random data node is chosen and all incoming points target that node until the node splits. Then, another random data node is chosen, and so on. In this experiment, we fix the dimensionality to 6 and we vary the node sizes. Table 3, demonstrates that m1 achieves a node storage utilization slightly above 50%, whereas, m2 achieves very good average data node storage utilization. Thus, the average number of accessed nodes per range query is considerably lower.

V. CONCLUSION

We proposed a new data node splitting method for the hB-tree or the KDB-tree. Since data nodes comprise the majority of nodes in a tree index, higher data node storage utilization can improve search performance. There is a need for indexes in medium dimensionality that can efficiently answer kNN queries.

splitting method	k	INS	DNS	nodes per tree level	utilization (%)	average nodes accessed
m1	2	5	10	1,2,8,36,173,730,3211,13946	71,71	4,64
m2	2	5	10	1,2,8,37,160,720,3118,13606	73,5	4,65
m1	3	5	10	1,2,9,42,173,743,3212,13956	71,65	9,72
m2	3	5	10	1,2,9,40,166,719,3157,13591	73,58	9,39
m1	10	5	10	1,2,8,33,163,757,3284,13929	71,79	877,38
m2	10	5	10	1,2,8,33,153,701,3144,13599	73,53	860,81
m1	15	5	10	1,2,8,39,184,808,3396,14171	70,57	14005,87
m2	15	5	10	1,2,9,39,178,794,3364,14183	70,51	13859,64

Table 2: Node storage utilization and query efficiency per splitting method for uniform data and insertion pattern and varying dimensionality

splitting method	k	INS	DNS	nodes per tree level	utilization (%)	average nodes accessed
m1	6	5	10	1,4,15,57,231,974,3954,16666	54,55	1254,37
m2	6	5	10	1,2,8,35,151,640,2821,12352	73,6	1038,52
m1	6	25	50	1,15,219,3920	51,03	447,47
m2	6	25	50	1,8,156,2841	70,41	372,11
m1	6	50	100	1,2,62,1980	50,52	284,84
m2	6	50	100	1,43,1422	70,34	243,26

Table 3: Node storage utilization per splitting method for uniform data, block insertion pattern and varying node sizes

So, we examined whether our splitting method improves the performance of the above mentioned indexes.

We defined the notion of paired data nodes, and we used this notion to propose the new splitting method. Our experiments show that redistribution works really well and improves data node storage utilization and range query performance.

REFERENCES

[Beckmann, N., Kriegel, H.-P., Schneider, R., and Seeger, B. \(1990\). The r*-tree: an efficient and robust access method for points and rectangles. In Proceedings of the 1990 ACM SIGMOD international conference on Management of data, SIGMOD '90, pages 322–331, New York, NY, USA. ACM.](#)

[Berchtold, S., Bohm, C., and Kriegel, H.-P. \(1998\). The pyramid-technique: towards breaking the curse of dimensionality. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, SIGMOD '98, pages 142–153, New York, NY, USA. ACM.](#)

[Comer, D. \(1979\). Ubiquitous b-tree. ACM Comput. Surv., 11:121–137.](#)

[Evangelidis, G., Lomet, D., and Salzberg, B. \(1997\). The hb⁷-tree: a multi-attribute index supporting concurrency, recovery and node consolidation. The VLDB Journal, 6:1–25.](#)

[Knuth, D. E. \(1973\). The Art of Computer Program-](#)

[ming, Vol 3, Sorting and Searching. Addison-Wesley Publ. Co., Reading, MA, USA.](#)

[Lomet, D. B. and Salzberg, B. \(1990\). The hb-tree: a multiattribute indexing method with good guaranteed performance. ACM Trans. Database Syst., 15:625–658.](#)

[Outsios, E. and Evangelidis, G. \(2010\). Achieving optimal average data node storage utilization in k-dimensional point data indexes. In Proceedings of the 5th International Scientific Conference, eRA: The Contribution of Information Technology to Science, Economy, Society and Education, Piraeus, Greece.](#)

[Robinson, J. T. \(1981\). The k-d-b-tree: a search structure for large multidimensional dynamic indexes. In Proceedings of the 1981 ACM SIGMOD international conference on Management of data, SIGMOD '81, pages 10–18, New York, NY, USA. ACM.](#)

[Yao, A. C.-C. \(1978\). On random 23 trees. Acta Informatica, 9:159–170. 10.1007/BF00289075.](#)

[Zhou, P. and Salzberg, B. \(2008\). The hb-pi* tree: An optimized comprehensive access method for frequent-update multi-dimensional point data. In Proceedings of the 20th international conference on Scientific and Statistical Database Management, SSDBM '08, pages 331–347, Berlin, Heidelberg. Springer-Verlag.](#)