

# Preface

We are delighted to see this edited book as the result of our intensive work over the past year. We succeeded in attracting high-quality submissions of which we could only include fourteen papers in this edited book. The present text aims at helping the readers both researchers from academia and practitioners from industry to grasp the basic concepts of information integrity and reusability which are very essential in this rapidly growing information era. Over the past decades, huge volumes of data have been produced and stored, large number of software systems have been developed and successfully utilized, and various products have been manufactured. Not all these developments could have been achieved without the investment of money, time, effort, and other resources. Over time, new data sources evolve and data integration continue to be an essential and vital requirement. Further, systems and products should be revised to adapt new technologies and to satisfy new needs. Instead of building from scratch, researchers in the academia and industry have realized the benefits of and concentrated on building new software systems by reusing some of the components that already exist and have been well tested. This trend avoids reinventing the wheel, however, comes at the cost of finding out the best set of existing components to be utilized and how they should be integrated together and with the new nonexisting components which are to be developed from scratch. These are nontrivial tasks and have led to challenging research problems in the academia and industry. Some of these issues have been addressed in this book, which is intended to be a unique resource for researchers, developers, and practitioners. In addition, the book will cover the latest developments and discoveries related to information reuse and integration in the academia and in the industry. It contains high-quality research papers written by experts in the field. Some of them are extended versions of the best papers which were presented at IEEE International Conference on Information Reuse and Integration, which was held in Las Vegas in August 2011.

The first paper “Mediators, Concepts and Practice” by Gio Wiederhold studies mediators, their concepts, and practice. Mediators are intermediary modules in large-scale information systems that link multiple sources of information to applications. They provide a means for integrating the application of encoded

knowledge into information systems. Mediated systems compose autonomous data and information services, permitting growth, and enable their survival in a semantically diverse and rapidly changing world. Constraints of scope are placed on mediators to assure effective and maintainable composed systems. Modularity in mediated architectures is not only a goal but also enables the goal to be reached. Mediators focus on semantic matching, while middleware provides the essential syntactic and formatting interfaces.

The second paper “A Combination Framework for Exploiting the Symbiotic Aspects of Process and Operational Data in Business Process Optimization” by Sylvia Radeschütz, Holger Schwarz, Marko Vrhovnik, and Bernhard Mitschang addresses the optimizing problem of a company’s business process. A profound analysis of all relevant business data in a company is necessary for optimizing business processes effectively. Current analyses typically run either on business process execution data or on operational business data. Correlations among the separate datasets have to be found manually under big effort. However, to achieve a more informative analysis and to fully optimize a company’s business, an efficient consolidation of all major data sources is indispensable. Recent matching algorithms are insufficient for this task since they are restricted either to schema or to process matching. They present a new matching framework to (semi)automatically combine process data models and operational data models for performing such a profound business analysis. They describe the algorithms and basic matching rules underlying this approach as well as an experimental study that shows the achieved high recall and precision.

The third paper “Efficient Range Query Processing on Complicated Uncertain Data” by Andrew Knight, Qi Yu, and Manjeet Rege investigates a special type of query, range queries, on uncertain data. Uncertain data has emerged as a key data type in many applications. New and efficient query processing techniques need to be developed due to the inherent complexity of this new type of data. They propose a threshold interval indexing structure that aims to balance different time-consuming factors to achieve an optimal overall query performance. They also present a more efficient version of their proposed structure which loads its primary tree into memory for faster processing. Experimental results are presented to justify the efficiency of the proposed query processing technique.

The fourth paper “Invariant Object Representation Based on Inverse Pyramidal Decomposition and Modified Mellin-Fourier Transform” by R. Kountchev, S. Rubin, M. Milanova, and R. Kountcheva presents a new method for invariant object representation based on the Inverse Pyramidal Decomposition (IPD) and modified Mellin-Fourier Transform (MFT). The so-prepared object representation is invariant against 2D rotation, scaling, and translation (RST). The representation is additionally made invariant to significant contrast and illumination changes. The method is aimed at content-based object retrieval in large databases. The experimental results obtained using the software implementation of the method proved its efficiency. The method is suitable for various applications, such as detection of children sexual abuse in multimedia files, search of handwritten and printed documents, and 3D objects, and represented by multi-view 2D images.

The fifth paper “Model Checking State Machines Using Object Diagrams” by Thouraya Bouabana-Tebibel discusses that UML behavioral diagrams are often formalized by transformation into a state-transition language that sets on a rigorously defined semantics. The state-transition models are afterwards model-checked to prove the correctness of the models construction as well as their faithfulness with the user requirements. The model checking is performed on a reachability graph, generated from the behavioral models, whose size depends on the models’ structure and their initial marking. The purpose of this paper is twofold. The author first proposes an approach to initialize formal models at any time of the system life cycle using UML diagrams. The formal models are Object Petri nets, OPNs for short, derived from UML state machines. The OPNs marking is mainly deduced from the sequence diagrams. Secondly, an approach is proposed to specify the association ends on the OPNs in order to allow their validation by means of OCL invariants. A case study is given to illustrate the approach throughout the paper.

The sixth paper “Measuring Stability of Feature Selection Techniques on Real-World Software Datasets” by Huanjing Wang, Taghi M. Khoshgoftaar, and Randall Wald studies the stability of different feature selection techniques on software data repositories. In the practice of software quality estimation, superfluous software metrics often exist in data repositories. In other words, not all collected software metrics are useful or make equal contributions to software defect prediction. Selecting a subset of features that are most relevant to the class attribute is necessary and may result in better prediction. This process is called feature selection. However, the addition or removal of instances can alter the subsets chosen by a feature selection technique, rendering the previously selected feature sets invalid. Thus, the robustness (e.g., stability) of feature selection techniques must be studied to examine the sensitivity of these techniques to changes in their input data (the addition or removal of instances). In this paper, authors test the stability of eighteen feature selection techniques as the magnitude of change to the datasets, and the size of the selected feature subsets are varied. All experiments were conducted on 16 datasets from three real-world software projects. The experimental results demonstrate that gain ratio shows the least stability while two different versions of ReliefF show the most stability, followed by the PRC- and AUC-based threshold-based feature selection techniques. Results also show that making smaller changes to the datasets has less impact on the stability of feature ranking techniques applied to those datasets.

The seventh paper “Analysis and Design: Towards Large-Scale Reuse and Integration of Web User Interface Components” by Hao Han, Peng Gao, Yinxing Xue, Chuanqi Tao, and Keizo Oyama studies the reuse and integration of Web user interface components. With the trend for Web information/functionality integration, application integration at the presentation and logic layers is becoming a popular issue. In the absence of OpenWeb service application programming interfaces, the integration of conventional Web applications is usually based on the reuse of user interface (UI) components, which partially represent the interactive functionalities of applications. In this paper, they describe some common problems of the current Web-UIComponent-based reuse and integration and propose a solution: a security-

enhanced “component retrieval and integration description” method. They also discuss the related technologies such as testing, maintenance, and copyright. Their purpose is to construct a reliable large-scale reuse and integration system for Web applications.

The eighth paper “Which Ranking for Effective Keyword Search Query over RDF Graphs?” by Roberto De Virgilio presents a theoretical study of YAANII, a novel technique to keyword-based search over semantic data. Ranking solutions is an important issue in information retrieval because it greatly influences the quality of results. In this context, keyword-based search approaches use to consider solutions sorting as least step of the overall process. Ranking and building solutions are completely separate steps running autonomously. This may penalize the retrieving information process because it binds to order all found matching elements including (possible) irrelevant information. The proposed approach presents a joint use of scoring functions and solution building algorithms to get the best results. The author demonstrates how effectiveness of the answers depends not so much on the quality of the scoring metrics but on the way such criteria are involved. Finally it is shown how YAANII overcomes other systems in terms of efficiency and effectiveness.

The ninth paper “ReadFast: Structural Information Retrieval from Biomedical Big Text by Natural Language Processing” by Michael Gubanov, Linda Shapiro, and Anna Pyayt discusses methods for retrieving information from large-scale text datasets. While the problem to find needed information on the Web is being solved by the major search engines, access to the information in Big text, large-scale text datasets, and documents (biomedical literature, e-books, conference proceedings, etc.) is still very rudimentary. Thus, keyword-search is often the only way to find the needle in the haystack. There is abundance of relevant research results in the Semantic Web research community that offers more robust access interfaces compared to keyword-search. Here authors describe a new information retrieval engine that offers advanced user experience combining keyword-search with navigation over an automatically inferred hierarchical document index. The internal representation of the browsing index as a collection of UFOs yields more relevant search results and improves user experience.

The tenth paper “Multiple Criteria Decision Support for Software Reuse: An Industrial Case Study” by Alejandra Yopez Lopez and Nan Niu reports a case study that applied SMART (Simple Multi-Attribute Rating Technique) to a company that considered reuse as an option of reengineering its Web site. In practice, many factors must be considered and balanced when making software reuse decisions. However, few empirical studies exist that leverage practical techniques to support decision-making in software reuse. The company’s reuse goal was set to maximize benefits and to minimize costs. They applied SMART in two iterations for the company’s software reuse project. The main difference is that the first iteration used the COCOMO (CONstructive COSt MODEL) to quantify the cost in the beginning of the software project. In the second iteration, they refined the cost estimation by using the COCOMO II model. This combined approach illustrates the importance of updating and refining the decision support for software reuse. The company was informed the optimal reuse percentage for the project, which was reusing

76–100 % of the existing artifacts and knowledge. This study not only shows that SMART is a valuable and practical technique that can be readily incorporated into an organization's software reuse program but also offers concrete insights into applying SMART in an industrial setting.

The eleventh paper "Using Local Principal Components to Explore Relationships Between Heterogeneous Omics Datasets" by Noor Alaydie and Farshad Fotouhi analyzes the relationships between a pair of data sources based on their correlation. In the post-genomic era, high-throughput technologies lead to the generation of large amounts of "omics" data such as transcriptomics, metabolomics, proteomics that are measured on the same set of samples. The development of methods that are capable to perform joint analysis of multiple datasets from different technology platforms to unravel the relationships between different biological functional levels becomes crucial. A common way to analyze the relationships between a pair of data sources based on their correlation is canonical correlation analysis (CCA). CCA seeks for linear combinations of all the variables from each dataset which maximize the correlation between them. However, in high-dimensional datasets, where the number of variables exceeds the number of experimental units, CCA may not lead to meaningful information. Moreover, when colinearity exists in one or both the datasets, CCA may not be applicable. In this paper, the authors present a novel method to extract common features from a pair of data sources using Local Principal Components and Kendall's Ranking (LPC-KR). The results show that the proposed algorithm outperforms CCA in many scenarios and is more robust to noisy data. Moreover, meaningful results are obtained using the proposed algorithm when the number of variables exceeds the number of experimental units.

The twelfth paper "Towards Collaborative Forensics" by Mike Mabey and Gail-Joon Ahn proposes a comprehensive framework to address the efficacious deficiencies of current practices in digital forensics. This framework, called Collaborative Forensic Framework (CUFF), provides scalable forensic services for practitioners who are from different organizations and have diverse forensic skills. In other words, this framework helps forensic practitioners collaborate with each other, instead of learning and struggling with new forensic techniques. In addition, the fundamental building blocks for the proposed framework and corresponding system requirements are described.

The thirteenth paper "From Increased Availability to Increased Productivity: How Researchers Benefit from Online Resources" by Joe Strathern, Samer Awadh, Samir Chokshi, Omar Addam, Omar Zarour, M. Ozair Shafiq, Orkun Öztürk, Omair Shafiq, Jamal Jida, and Reda Alhaji presents researchers with an analysis that summarizes data collection and evolution features in the World Wide Web. Authors have reviewed a large corpus of published work to extract the research features supported by advanced Web features, i.e., blogs, microblogging services, video-on-demand Web sites, and social networks. They have presented summary of their review and analysis. This will help in further analysis of the evolution of communities using and supporting the features of the continuously evolving World Wide Web.

The fourteenth paper “Integration of Semantics Information and Clustering in Binary-class Classification for Handling Imbalanced Multimedia Data” by Chao Chen and Mei-Ling Shyu proposes a novel binary-class classification framework that integrates the video semantics information and the clustering technique to address the data imbalance issue which is a major challenge in the classification task. Experiments are conducted to compare the proposed framework with other techniques that are commonly used to learn from imbalanced datasets. The experimental results on some highly imbalanced video datasets demonstrate that the proposed classification framework outperforms these comparative classification approaches about 3 % to 16 %.

Last but not the least, we would like to mention the hard workers behind the scene who have significant unseen contributions to the successful task that produced this valuable source of knowledge. We would like to thank the authors who submitted papers and the reviewers who provided detailed constructive reports which improved the quality of the papers. Various people from Springer deserve large credit for their help and support in all the issues related to publishing this book.

Ankara, Turkey  
Ontario, Canada  
Ankara, Turkey  
Houston, Texas

Tansel Özyer  
Keivan Kianmehr  
Mehmet Tan  
Jia Zeng

Information Reuse and Integration in Academia and  
Industry

Özyer, T.; Kianmehr, K.; Tan, M.; Zeng, J. (Eds.)

2013, XII, 306 p. 103 illus., Hardcover

ISBN: 978-3-7091-1537-4