Lecture Notes in Computer Science 6380

Abdelkader Hameurlain Josef Küng
Roland Wagner Torben Bach Pedersen
A Min Tjoa (Eds.)

Transactions on Large-Scale Data- and Knowledge-Centered Systems II

Springer

Editors-in-Chief

Abdelkader Hameurlain
Paul Sabatier University
Institut de Recherche en Informatique de Toulouse (IRIT)
118, route de Narbonne, 31062 Toulouse Cedex, France
E-mail: hameur@irit.fr

Josef Küng
Roland Wagner
University of Linz, FAW
Altenbergerstraße 69
4040 Linz, Austria
E-mail: {jkueng,rrwagner}@faw.at

Guest Editors

Torben Bach Pedersen
Aalborg University
Department of Computer Science
Selma Lagerløfs Vej 300
9220 Aalborg, Denmark
E-mail: tbp@cs.aau.dk

A Min Tjoa
Vienna University of Technology
Institute of Software Technology
Favoritenstr. 9-11/188
1040 Vienna, Austria
E-mail: amin@ifs.tuwien.ac.at

Library of Congress Control Number: 2009932361

CR Subject Classification (1998): H.2, I.2.4, H.3, H.4, J.1, H.2.8

ISSN 0302-9743 (Lecture Notes in Computer Science)
ISSN 1869-1994 (Trans. on Large-Scale Data- and Knowledge-Centered Systems)
ISBN-10 3-642-16174-X Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16174-2 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

This special issue of TLDKS contains two kinds of papers. First, it contains a selection of the best papers from the 11th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2009), which was held from August 31 to September 2, 2009 in Linz, Austria. Second, it contains a special section of papers on a particularly challenging domain in information retrieval, namely patent retrieval.

Over the last decade, the International Conference on Data Warehousing and Knowledge Discovery (DaWaK) has established itself as one of the most important international scientific events within data warehousing and knowledge discovery. DaWaK brings together a wide range of researchers and practitioners working on these topics. The DaWaK conference series thus serves as a leading forum for discussing novel research results and experiences within the field. The 11th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2009) continued the tradition by disseminating and discussing innovative models, methods, algorithms, and solutions to the challenges faced by data warehousing and knowledge discovery technologies.

The papers presented at DaWaK 2009 covered a wide range of issues within data warehousing and knowledge discovery. Within data warehousing and analytical processing, the topics covered were: data warehouse modeling including advanced issues such as spatio-temporal warehouses and DW security, OLAP on data streams, physical design of data warehouses, storage and query processing for data cubes, advanced analytics functionality, and OLAP recommendation. Within knowledge discovery and data mining, the topics included: stream mining, pattern mining for advanced types of patterns, advanced rule mining issues, advanced clustering techniques, spatio-temporal data mining, data mining applications, as well as a number of advanced data mining techniques. It was encouraging to see that many papers covered important emerging issues such as: spatio-temporal data, streaming data, non-standard pattern types, advanced types of data cubes, complex analytical functionality including recommendations, multimedia data, mssing and noisy data, as well as real-world applications within genes, and the clothing and telecom industries. The wide range of topics bears witness to the fact that the data warehousing and knowledge discovery field is dynamically responding to the new challenges posed by novel types of data and applications.

The DaWaK 2009 Call for Papers attracted a large number of quality submissions. From 124 submitted abstracts, we received 100 papers from 17 countries in Europe, North America, and Asia. The program committee finally selected 36 papers, yielding an acceptance rate of 36%. The DaWaK proceedings were published by Springer in the LNCS series.

A few of the papers were selected by the DaWaK PC chairs based on both quality and potential, and the authors were invited to submit significantly extended versions for a new round of reviewing. After a thorough refereeing process, including further revisions of the papers, the following three papers were finally accepted for inclusion in this special issue of TLDKS.

In the paper “Fast Loads and Fast Queries“, Graefe and Kuno aim at the double goal of achieving fast bulk loading, while still supporting fast queries through computing redundant search structures on the fly. The paper introduces a range of new techniques to achieve this, named zone filters, zone indexes, adaptive merging, and partition filters, respectively. These techniques address the limitations of earlier techniques like Netezza zone maps without reducing the advantages. The proposed data structures can be created on the fly, as a “side effect” of the load process. All required analyses can be performed with a moderate amount of new data in the main memory buffer pool, and traditional sorting and indexing are not required. However, the observed query performance is as good as that of zone maps where those can be used, and is better than that of zone maps for the (many) predicates where zone maps are ineffective. Finally, simulations show that the query performance is comparable to query processing in a database with traditional indexing, but without the long loading time that such indexing requires.

In the paper “Discovery of Frequent Patterns in Transactional Data Streams”, Ng and Dash investigate the problem of discovering frequent patterns in a transaction stream. They first survey two common methods from the literature: 1) approximate counting, e.g., lossy counting (LCA), using a lower support threshold and 2) maintaining a running sample, e.g., reservoir sampling (Algo-Z), and generating frequent itemsets on demand from this sample. Then, the pros and cons of each method are discussed. The authors then propose a novel sampling algorithm, called DSS, which selects a transaction to include in the sample based on single item histograms. Finally, an experimental comparison between the three approaches is performed, showing that DSS is the most accurate, then LCA, then Algo-Z, while Algo-Z is the fastest, followed by DSS, and then LCA.

In the paper “Efficient Online Aggregates in Dense-Region-Based Data Cube Representations”, Haddadin and Lauer investigate the space- and time-efficient in-memory representation of large data cubes. The proposed solution builds on the well known idea of identifying dense sub-regions of the cube and storing dense and sparse regions separately, but improves upon it by focusing not only on space- but also on time-efficiency. This is done both for the initial dense-region extraction and for queries carried out in the resulting hybrid data structure. The paper first describes a pre-aggregation method for representing dense sub-cubes which supports efficient online aggregate queries as well as cell updates. Then, the sub-cube extraction approach is presented. Two optimization methods that trade available memory for increased aggregate query performance are presented, along with the adaptation to multi-core/multi-processor architectures. Experiments with several real-world data are performed, showing how the algorithms can be tuned by setting a number of parameters.

Moving from the analysis of structured data to a semi-structured domain, patent retrieval poses a range of demanding challenges from an information retrieval perspective. Part of this is due to the sheer volume of data, part to the rather specific requirements on the retrieval scenarios. Patent retrieval, for example, is one of the few settings where recall is commonly more important than precision. Different to normal web retrieval, where the goal is to provide a set of relevant documents within the top-10 documents retrieved, patent retrieval requires all documents to be found, with result lists being scanned ranging well into the hundreds.

The two papers selected for this special section address a specific characteristic of this domain, namely the question of whether a system is able to find specific documents at all. As only a limited number of documents can be examined in any setting, there may be documents that do not appear within the top-n ranks for any realistic query, thus remaining practically irretrievable. Both papers in this special section address this retrievability challenge from two different perspectives.

The paper “Improving Access to Large Patent Corpora” by Bache and Azzopardi analyzes the characteristics of the retrieval system itself, measuring the system bias towards specific document characteristics. Specifically, a system’s sensitivity to term frequency characteristics of the documents, length variations and convexity, i.e., managing the impact of phrases, are evaluated. It then proposes a hybrid retrieval system, combining both exact as well as best-match principles.

The paper “Improving Retrievability and Recall by Automatic Corpus Partitioning” by Bashir and Rauber aims to overcome the retrievability challenge by automatically classifying documents into ones with potentially high and low retrievability based on statistical characteristics of word distributions. Using a split corpus, separate retrieval engines can be applied, allowing documents from the lower retrievability section of a document corpus to find their way into the result set, thus ensuring better coverage and mitigating the risk of having un-findable documents.

As the Guest Editors of this special issue, we would like to thank all the referees, both the DaWaK 2009 PC members and the extra reviewers for the special issue, for their careful and dedicated work. We hope you will enjoy the papers that follow and see them as bearing witness to the high quality of the DaWaK conference series as well as to the specific challenges faced by retrieval engines in highly specialized domains of large volume data.

July 2010

Torben Bach Pedersen
A Min Tjoa

Editorial Board

Table of Contents

Data Warehousing and Knowledge Discovery

Information Retrieval