Lecture Notes in Computer Science 9940

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

Editorial Board

More information about this series at http://www.springer.com/series/8637

Abdelkader Hameurlain · Josef Küng
Roland Wagner · Qimin Chen (Eds.)

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

Special Issue on Database- and Expert-Systems Applications

Springer

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-662-53454-0 ISBN 978-3-662-53455-7 (eBook)
DOI 10.1007/978-3-662-53455-7

Library of Congress Control Number: 2015943846

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Berlin Heidelberg

Preface

The 26th International Conference on Database and Expert Systems Applications, DEXA 2015, held in Valencia, Spain, September 1–4, 2015, provided a premier forum and unique opportunity for researchers, developers, and users from different disciplines to present the state of the art, exchange research ideas, share industry experiences, and explore future directions at the intersection of data management, knowledge engineering, and artificial intelligence. This special issue of Springer's *Transactions on Large-Scale Data- and Knowledge-Centered Systems (TLDKS)* contains extended versions of selected papers presented at the conference. While these articles describe the technical trend and the breakthroughs made in the field, the general message delivered from them is that turning big data to big value requires incorporating cutting-edge hardware, software, algorithms and machine-intelligence.

Efficient graph-processing is a pressing demand in social-network analytics. A solution to the challenge of leveraging modern hardware in order to speed up the similarity join in graph processing is given in the article "Accelerating Set Similarity Joins Using GPUs", authored by Mateus S. H. Cruz, Yusuke Kozawa, Toshiyuki Amagasa, and Hiroyuki Kitagawa. In this paper, the authors propose a GPU (Graphics Processing Unit) supported set similarity joins scheme. It takes advantage of the massive parallel processing offered by GPUs, as well as the space efficiency of the MinHash algorithm in estimating set similarity, to achieve high performance without sacrificing accuracy. The experimental results show more than two orders of magnitude performance gain compared with the serial version of CPU implementation, and 25 times performance gain compared with the parallel version of CPU implementation. This solution can be applied to a variety of applications such as data integration and plagiarism detection.

Parallel processing is the key to accelerating machine-learning on big data. However, many machine leaning algorithms involve iterations that are hard to be parallelized from either the load balancing among processors, memory access overhead, or race conditions, such as those relying on hierarchical parameter estimation. The article "Divide-and-Conquer Parallelism for Learning Mixture Models", authored by Takaya Kawakatsu, Akira Kinoshita, Atsuhiro Takasu, and Jun Adachi, addresses this problem. In this paper, the authors propose a recursive divide-and-conquer-based parallelization method for high-speed machine learning, which uses a tree structure for recursive tasks to enable effective load balancing and to avoid race conditions in memory access. The experiment results show that applying this mechanism to machine learning can reach a scalability superior to FIFO scheduling, with robust load imbalance.

Maintaining multistore systems has become a new trend for integrated access to multiple, heterogeneous data, either structured or unstructured. A typical solution is to extend a relational query engine to use SQL-like queries to retrieve data from other data sources such as HDFS, which, however, requires the system to provide a relational view of the unstructured data. An alternative approach is proposed in the article "Multistore Big Data Integration with CloudMdsQL", authored by Carlyna

Bondiombouy, Boyan Kolev, Oleksandra Levchenko, and Patrick Valduriez. In this paper, a functional SQL-like query language (based on CloudMdsQL) is introduced for integrated data retrieved from different data stores, therefore taking full advantage of the functionality of the underlying data management frameworks. It allows user defined map/filter/reduce operators to be embedded in traditional SQL statements. It further allows the filtering conditions to be pushed down to the underlying data processing framework as early as possible for the purpose of optimization. The usability of this query language and the benefits of the query optimization mechanism are demonstrated by the experimental results.

One of the primary goals of exploring big data is to discover useful patterns and concepts. There exist several kinds of conventional pattern matching algorithms; for instance, the terminology-based algorithms are used to compare concepts based on their names or descriptions, the structure-based algorithms are used to align concept hierarchies to find similarities; the statistic-based algorithms classify concepts in terms of various generative models. In the article "Ontology Matching with Knowledge Rules", authored by Shangpu Jiang, Daniel Lowd, Sabin Kafle, and Dejing Dou, the focus is shifted to aligning concepts by comparing their relationships with other known concepts. Such relationships are expressed in various ways – Bayesian networks, decision trees, association rules, etc.

The article "Regularized Cost-Model Oblivious Database Tuning with Reinforcement Learning", authored by Debabrota Basu, Qian Lin, Weidong Chen, Hoang Tam Vo, Zihong Yuan, Pierre Senellart, and Stephane Bressan, proposes a machine learning approach for adaptive database performance tuning, a critical issue for efficient information management, especially in the big data context. With this approach, the cost model is learned through reinforcement learning. In the use case of index tuning, the executions of queries and updates are modeled as a Markov decision process, with states represented in database configurations, actions causing configuration changes, corresponding cost parameters, as well as query and update evaluations. Two important challenges in the reinforcement learning process are discussed: the unavailability of a cost model and the size of the state space. The solution to the first challenge is to learn the cost model iteratively, using regularization to avoid overfitting; the solution to the second challenge is to prune the state space intelligently. The proposed approach is empirically and comparatively evaluated on a standard OLTP dataset, which shows competitive advantage.

The article "Workload-Aware Self-tuning Histograms for the Semantic Web", authored by Katerina Zamani, Angelos Charalambidis, Stasinos Konstantopoulos, Nickolas Zoulis, and Effrosyni Mavroudi, further discusses how to optimize the histograms for semantic Web. As we know, query processing systems typically rely on histograms which represent approximate data distribution, to optimize query execution. Histograms can be constructed by scanning the datasets and aggregating the values of the selected fields, and progressively refined by analyzing query results. This article tackles the following issue: histograms are typically built from numerical data, but the Semantic Web is described with various data types which are not necessarily numeric. In this work a generalized histograms framework over arbitrary data types is established with the formalism for specifying value ranges corresponding to various datatypes. Then the Jaro-Winkler metric is introduced to define URI ranges based on the

hierarchical nature of URI strings. The empirical evaluation results, conducted using the open-sourced STRHist system that implements this approach, demonstrate its competitive advantage.

We would like to thank all the authors for their contributions to this special issue. We are grateful to the reviewers of these articles for their invaluable efforts in collaborating with the authors to deliver readers the precise ideas, theories, and solutions on the above state-of-the-art technologies. Our deep appreciation also goes to Prof. Roland Wagner, Chairman of the DEXA Organization, Ms. Gabriela Wagner, Secretary of DEXA, the distinguished keynote speakers, Program Committee members, and all presenters and attendees of DEXA 2015. Their contributions help to keep DEXA a distinguished platform for exchanging research ideas and exploring new directions, thus setting the stage for this special TLDKS issue.

June 2016

Qiming Chen
Abdelkader Hameurlain

Organization

Editorial Board

Reza Akbarinia	Inria, France
Bernd Amann	LIP6 - UPMC, France
Dagmar Auer	FAW, Austria
Stéphane Bressan	National University of Singapore, Singapore
Francesco Buccafurri	Università Mediterranea di Reggio Calabria, Italy
Qiming Chen	HP-Lab, USA
Mirel Cosulschi	University of Craiova, Romania
Dirk Draheim	University of Innsbruck, Austria
Johann Eder	Alpen Adria University Klagenfurt, Austria
Georg Gottlob	Oxford University, UK
Anastasios Gounaris	Aristotle University of Thessaloniki, Greece
Theo Härder	Technical University of Kaiserslautern, Germany
Andreas Herzig	IRIT, Paul Sabatier University, France
Dieter Kranzlmüller	Ludwig-Maximilians-Universität München, Germany
Philippe Lamarre	INSA Lyon, France
Lenka Lhotská	Technical University of Prague, Czech Republic
Vladimir Marik	Technical University of Prague, Czech Republic
Franck Morvan	Paul Sabatier University, IRIT, France
Kjetil Nørvåg	Norwegian University of Science and Technology, Norway
Gultekin Ozsoyoglu	Case Western Reserve University, USA
Themis Palpanas	Paris Descartes University, France
Torben Bach Pedersen	Aalborg University, Denmark
Günther Pernul	University of Regensburg, Germany
Sherif Sakr	University of New South Wales, Australia
Klaus-Dieter Schewe	University of Linz, Austria
A Min Tjoa	Vienna University of Technology, Austria
Chao Wang	Oak Ridge National Laboratory, USA

External Reviewers

Nadia Bennani	INSA of Lyon, France
Miroslav Bursa	Czech Technical University, Prague, Czech Republic
Eugene Chong	Oracale Incorporation, USA
Jérôme Darmont	University of Lyon, France
Flavius Frasincar	Erasmus University Rotterdam, The Netherlands
Jeff LeFevre	HP Enterprise, USA

Junqiang Liu	Zhejiang Gongshang University, China
Rui Liu	HP Enterprise, USA
Raj Sundermann	Georgia State University, USA
Lucia Vaira	University of Salento, Italy
Kevin Wilkinson	HP Enterprise, USA
Shaoyi Yin	Paul Sabatier University, Toulouse, France
Qiang Zhu	The University of Michigan, USA

Contents