SCALABLE HIGH PERFORMANCE COMPUTING FOR KNOWLEDGE DISCOVERY AND DATA MINING

edited by

Paul Stolorz

Jet Propulsion Laboratory

and

Ron Musick

Lawrence Livermore National Laboratory

A Special Issue of

**DATA MINING AND
KNOWLEDGE DISCOVERY**

Volume 1, No. 4 (1997)

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

DATA MINING AND KNOWLEDGE DISCOVERY

Volume 1, No. 4

Special Issue: Scalable High Performance Computing for KDD

Guest Editorial	<i>Paul Stolorz and Ron Musick</i>	1
-----------------------	------------------------------------	---

Full Paper Contributors

Parallel Algorithms for Discovery of Association Rules		
..... <i>Mohammed J. Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara and Wei Li</i>		5
A Distributed Algorithm for Content Based Indexing of Images by Projections on Ritz Primary Images	<i>Haim Schweitzer</i>	37
High Performance OLAP and Data Mining on Parallel Computers		
..... <i>Sanjay Goil and Alok Choudhary</i>		53
Halo World: Tools for Parallel Cluster Finding in Astrophysical <i>N</i> -Body Simulations.....	<i>David W. Pfizner, John K. Salmon and Thomas Sterling</i>	81

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

ISBN 978-1-4613-7595-1 ISBN 978-1-4615-5669-5 (eBook)
DOI 10.1007/978-1-4615-5669-5

Copyright © 1998 Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 1998

Softcover reprint of the hardcover 1st edition 1998

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Springer Science+Business Media, LLC .

Printed on acid-free paper.

Editorial

PAUL STOLORZ
Jet Propulsion Laboratory

pauls@aig.jpl.nasa.gov

RON MUSICK
Lawrence Livermore National Laboratory

rmusick@llnl.gov

It is by now a commonplace observation that our world is overwhelmed by the sheer volume and complexity of "information" available to us. Advances in data acquisition, storage and transmission technologies have greatly outstripped our ability to analyze and organize this information using the time-honored methods that have served us well in the past. There is a general feeling that without more automated and efficient approaches to data analysis and characterization, vast amounts of data will essentially be wasted.

The field of Knowledge Discovery in Databases (KDD) has arisen in an effort to marshal interdisciplinary resources and expertise to tackle this vexing state of affairs. By combining ideas drawn from fields such as databases, machine learning, statistics, visualization and parallel and distributed computing, its goal is to generate an integrated approach to knowledge discovery that is more powerful and richer than the sum of its parts.

There are a number of different working definitions of KDD. A common thread in all definitions is that large database size is a fundamental characteristic of KDD (see (Fayyad, Piatetsky-Shapiro and Smyth, 1996) for a comprehensive discussion of these definitional issues). The focus of this Special Issue of Data Mining and Knowledge Discovery is the development of powerful new ways of applying KDD methods on scalable high-performance computing platforms as one of the crucial ingredients needed to deal with large database size.

Of course, scalable platforms and implementations alone are no panacea for the efficiency problems involved with analyzing massive datasets. The underlying algorithms themselves, by nature, must also be scalable. Scalability means many things to different people. We can view a code to be scalable if it can effectively use additional computational resources to solve larger problems. More precisely, as we add system resources (e.g. processors) in proportion to increasing problem size, the total work, storage and communication per processor should not depend on the overall problem size. Furthermore, unless these scalable algorithms are fully integrated with powerful data management and storage systems, and enhanced with methods from machine learning, statistics and visualization, their usefulness will be extremely limited.

Accordingly, this special issue describes advances in scalable implementations of several important KDD techniques, drawn from a number of different domain areas. Each of the contributions considers a substantial existing KDD problem involving large amounts of data, describes a set of appropriate analysis techniques grounded in one or more of the relevant disciplines, and then implements and tests parallel versions of the techniques on real-world high-performance computing platforms. The implementations have been used to analyze, in some detail, the issues associated with scaling platform sizes to match growing data volumes and complexities. They highlight the great progress that can be made on KDD

problems by exploiting scalable infrastructure, while at the same time providing insight into the current limitations, and identifying the main challenges that must be met as the field matures.

We begin with a well-known data mining method with its roots in the database field, namely association rules. Zaki, Parthasarathy, Ogihara and Li describe a parallel method for discovering these rules, and implement this discovery process efficiently on scalable platforms. Association rules have been an early success story in the application of data mining notions to the extraction of patterns from relational databases. They have had a dramatic impact in business environments which have built or inherited substantial databases, and are clearly a natural and important target for implementation on high-performance platforms.

Another important question concerns the need to address large datasets by content. Schweitzer discusses a novel method for performing content based indexing of images in a distributed setting. This problem is important in data mining problems where searches for similarity are needed. With vast data volumes, we can typically no longer afford the luxury of matching two patterns in a dataset exactly. There may be many patterns with only slight differences that we want to group together as a single class, either as indexes for efficient access, or as an aid to modeling, or both. Rapid methods for this content-based access are crucial.

Goil and Choudhary tackle the OLAP domain, an exciting area that is becoming increasingly common in decision support systems. They describe a method for implementing the DataCube (Gray, et al., 1997), a relational operator designed to support searches for anomalies and unusual patterns. Parallelization is crucial here because the multidimensional data representations involved rapidly lead to huge computational demands.

Pfitzer, Salmon and Sterling describe an application of statistical clustering ideas to the task of analyzing astrophysical N-body simulations on scalable machines. Given the ubiquity of clustering in data mining approaches, their method, while initially developed for the N-body problem, should prove to be of great value for efficient parallel spatial clustering for a number of other KDD problems as well.

Future Challenges

The papers in this issue cover several applications relevant to KDD (see [3,4] for other typical academic and industrial applications). The work takes place on a variety of architectures, ranging from custom-built massively parallel distributed memory machines, to clusters of off-the-shelf components which can be rapidly assembled into dedicated parallel devices, to groups of workstations connected by local area networks. Although these architectures represent several of the most important configurations for KDD, they are clearly not the whole story. The future will undoubtedly feature exciting developments including coordinated data mining approaches across wide-area networks, e.g. "meta-supercomputing" approaches such as those pioneered by the CASA Gigabit Network initiative, and implementations on the ubiquitous internet.

Just as parallel database servers have now come into their own as important components of database solutions, a major challenge for KDD is to integrate data mining query primitives seamlessly with database management systems. Not only must important query primitives be identified and implemented efficiently, but the arcane details of parallel decomposition must also be made transparent to the user. The end user should be unable to tell whether

her query is being executed by a 256-node distributed memory machine or a single desktop device (except, of course, by the speed of response!).

We have in fact just begun to explore the computational needs of KDD algorithms, and of how scalable machines can meet them. Several of the important issues have been addressed in the contributions presented here, but there are many other aspects that must be dealt with if KDD is to live up to its potential. Areas that should be (and are being) explored include:

1. Visualization of large dimensional datasets 2. Parallel forms of unsupervised and supervised learning 3. Handling growth and change in data incrementally 4. Dealing with heterogeneous and widely distributed databases 5. Scalable I/O for data-rich problems with relatively small computational demands.

The contributions presented here each show the value of applying selected parallel techniques to specific KDD problems. However, their real value is to point the way to general implementations in the future that will be useful for any number of KDD tasks. For example, once a powerful spatial clustering method is parallelized efficiently, it can be applied to any data mining task that requires such a method. The articles in this special issue will be most useful if read in this light.

Scalable high-performance computers promise to have a huge impact on the KDD field. Without them, our claims that KDD is designed to address the problems inherent in "massive" datasets ring a little hollow. With their help, we are confident that great advances can be made in the quest to take full advantage of the mounds of data that exist and the mountains of data that are being created.

References

- Fayyad, U. Piatetsky-Shapiro, G. and Smyth, P. 1996. "From Data Mining to Knowledge Discovery: An Overview", in *Advances in Knowledge Discovery and Data Mining*, Fayyad et al, (Eds), MIT Press.
- Gray, J. Chaudhuri, S. Bosworth, A. Layman, A. Reichart, D. Venkatrao, M. Fellows, F. and Pirahesh, H. 1997. "DataCube: A Relational Aggregation Operator Generalizing Group-by, Cross-tab and Sub-totals", *Data Mining and Knowledge Discovery*, 1:29-53.
- Fayyad, U. Haussler, D. and Stolorz, P. 1996. "Mining Science Data". *Communications of the ACM* 39:51-57.
- Brachman, R. Khabaza, T. Kloesgen, W. Piatetsky-Shapiro, G. and Simoudis, E. 1996. "Mining Business Databases". *Communications of the ACM* 39:42-48.