Report from the Workshop on Distributed and Parallel Knowledge Discovery, ACM SIGKDD-2000

Hillol Kargupta
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250
hillol@cs.umbc.edu

Joydeep Ghosh
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712-1084
ghosh@ece.utexas.edu

Vipin Kumar
Department of Computer Science and Army High Performance Research Center, University of Minnesota
Minneapolis, MN 55455
kumar@cs.umn.edu

Zoran Obradovic
Center for Information Science and Technology
Temple University
Philadelphia PA 19122
zoran@joda.cis.temple.edu

ABSTRACT

This report presents a summary of the activities that took place in the 2000 ACM SIGKDD workshop on Distributed and Parallel Knowledge Discovery.

Keywords

Distributed and parallel knowledge discovery.

1. INTRODUCTION

Information is most useful when it can be transformed into actionable knowledge. Such transformations require efficient data access, analysis, and presentation of the outcome in a timely manner. The development of large distributed environments like the Internet and corporate intranets has added a new dimension to this process---a large number of distributed sources of data that can be used for discovering knowledge. In an increasingly mobile and wired world with a large number of distributed data sources, knowledge discovery (KDD) demands due attention to the cost of data communication for centralizing data. Limited network bandwidth, data security, and existing organizational structure of the applications environment often contribute to this communication cost. The field of distributed KDD studies algorithms, systems, and human-computer interaction issues for knowledge discovery applications in such distributed environments.

This document presents a summary of the events at the 2000 ACM SIGKDD workshop on distributed and parallel knowledge discovery. Although the workshop was primarily focused on distributed KDD, it encouraged participation of the high performance parallel KDD research community and hosted a special session for this area.

2. SUMMARY OF THE WORKSHOP

The workshop on distributed and parallel knowledge discovery held at the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) attracted both researchers and practitioners. The workshop was focused on the state-of-the-art distributed and parallel knowledge discovery algorithms, systems, and application related issues. Approximately forty participants attended the workshop. The workshop had fourteen presentations, including two invited talks.

The workshop started with an invited talk by Michael Heytens on high performance knowledge discovery in a zero-latency enterprise (ZLE) environment at Compaq . This talk discussed Compaq's ZLE system architecture and the major KDD-related challenges faced by this project.

This was followed by a general session chaired by Mohammed Zaki from Rensselaer Polytechnic. Turinsky and Grossman presented a framework to study the trade-off between completely distributed and centralized distributed data mining (DDM). It provided an analytical basis for capturing the cost of communication and data analysis in order to optimally design a DDM application. McClean, Scotney, and Greer presented a novel maximum likelihood estimation-based technique for clustering distributed heterogeneous databases. Kargupta, Huang, Sivakumar, and Johnson presented a new technique for performing distributed principal component analysis (PCA) from heterogeneous data. They also explained how the distributed PCA algorithm can be used for clustering distributed financial data. Sayal and Scheuermann presented another distributed clustering algorithm for mining web-access data. Their presentation was followed by a talk on cost complexity-based pruning of ensemble classifiers by Prodromidis and Stolfo.

The afternoon session was kicked-off by an invited talk from Laveen Kanal, of LNK Corporation and University of Maryland, one of the pioneers in pattern recognition. He presented a historical perspective of the field of pattern recognition, its relation with distributed data mining, and some of his recent

work in this area. This was followed by a general session chaired by Srinivasan Parthasarathy from Ohio State University. This session was mainly dedicated for high performance parallel data mining. Kamath and Cant'u-Paz described a new tool-box for developing parallel object-oriented data mining system. They also discussed some of the scientific data mining applications that they are considering. Wang and Skillicorn presented a paper on parallel inductive logic for data mining. Dash and Liu presented a fast and memory-efficient approach to density-based clustering.

The next session had four work-in-progress papers. It was chaired by Zoran Obradovic, Temple University. Forman and Zhang presented their work on parallel non-approximate recasting of center-based clustering algorithms. Hall, Bowyer, Kegelmeyer, Moore, and Chao presented their DDM experiments on very large data sets. Caragea, Silvescu, and Honavar presented a theoretical framework for analysis and synthesis of incremental and distributed learning agents. Kimm and Ryu presented a CORBA-based framework for DDM over heterogeneous networks.

The concluding session was the panel discussion. Bob Grossman (University of Illinois at Chicago), Chandrika Kamath (Lawrence Livermore National Laboratory) and Hillol Kargupta (Washington State University) discussed the advances and future directions of the field of distributed and parallel knowledge discovery.

The organizers sincerely hope that the workshop created a stimulating environment for further growth of the field of distributed and parallel KDD. Selected papers from the workshop will be published in a special issue of the Knowledge and Information Systems Journal, Springer-Verlag in 2001. Further details about the developments can be found at http://www.distributed-kdd.com.

3. ACKNOWLEDGMENTS

The authors would like to thank Byung-Hoon Park for the publicity and local management of the workshop.

About the authors:

Hillol Kargupta is an Assistant Professor at the Computer Science and Electrical Engineering department of University of Maryland, Baltimore County. He earned his PhD. in Computer Science from the University of Illinois at Urbana-Champaign. His research is supported by a National Science Foundation (NSF) Career award, several NSF grants, the American Cancer Society, and other organizations. His other accomplishments include a Los Alamos Laboratory outstanding technical achievement award, the 1996 SIAM annual best student paper award, and more than fifty peer reviewed publications. He is also the co-editor of the book entitled Advances in Distributed and Parallel Knowledge Discovery, MIT/AAAI Press.

Joydeep Ghosh is a Professor and Archie Straiton Fellow at the Department of Electrical and Computer Engineering of the University of Texas at Austin. He received his PhD. from University of Southern California. He received the 1992 Darlington Award given by the IEEE Circuits and Systems Society for the Best Paper in the areas of CAS/CAD. He has published more than one hundred and fifty refereed papers and co-edited six books. He is an associate editor for IEEE Trans. Neural Networks, Pattern Recognition, Journal of Smart Engineering Design, and Neural Computing Surveys .

Vipin Kumar is a Professor of Computer Science and the Director of the Army High Performance Computing Research Center, University of Minnesota. He earned his PhD. degree in Computer Science from University of Maryland, College Park. He has authored over one hundred research articles, and co-edited or co-authored five books including the widely used text book Introduction to Parallel Computing. He serves on the editorial boards of IEEE Concurrency, Parallel Computing, the Journal of Parallel and Distributed Computing. He is a Fellow of IEEE.

Zoran Obradovic is the Director of the Center for Information Science and Technology, Temple University. His research interests focus on solving challenging Bioinformatics, E-Commerce and Computational Finance problems by developing and integrating data mining and statistical learning technology for an efficient knowledge discovery at large sequence, spatial and temporal databases. His research is Funded by NSF, NIH, DOE and industry, during the last decade he contributed to about ninety refereed articles on these and related topics. He is an editorial board member at Multiple Valued Logic, Journal of Computational Intelligence in Finance and IEEE Transaction on Education.