Preview

Clustering, or partitioning of datasets into subsets (also called clusters) so that the members of a cluster are more similar to each other than to members of other clusters is a long standing problem with rich history of documented research. Explosive use of the Internet and recent advances in information technology bring new challenges to this exciting research area. In many important applications the data resides in high dimensional vector spaces. While processing of high dimensional data represents computational challenges, in many cases the high dimensional vectors are sparse. The sparsity allows an efficient data clustering. In this one day workshop papers will be presented by experts from academia and industry. Clustering issues that will be covered include: bioinformatics applications, nearest neighborhood techniques, high dimensional clustering, clustering in low dimensional spaces. The workshop held in with the Second SIAM Conference on Data Mining brings together applied mathematicians, computers scientists, and computational statisticians working toward design of next generation clustering algorithms and software.

Inderjit S. Dhillon
Department of Computers Science
University of Texas

Jacob Kogan
Department of Mathematics and Statistics
University of Maryland Baltimore County

Acknowledgements

Special thanks to the members of the Program Committee for their diligent efforts in reviewing all the manuscripts submitted.

Program Committee

Cliff Behrens, Telcordia Technologies
Paul Bradley, digiMine Inc.
Dan Boley, University of Minnesota
Kui-Yu Chang, Interwoven Inc., Austin
Ming Gu, University of California, Berkeley
George Karypis, University of Minnesota
Jon Kettenring, Telcordia Technologies
Shailesh Kumar, HNC
Edward Marcotte, University of Texas
Dharmendra Modha, IBM Almaden Research Center
Ray Mooney, University of Texas, Austin
Nick Street, University of Iowa
Mark Teboulle, Tel-Aviv University

Workshop Schedule (presenters are boldfaced)

Session I: Clustering in Bioinformatics

9:00 - 9:45 Plenary Speaker: Prof. Edward Marcotte

9:45 - 10:15 Block Clustering on Continuous Data, Gérard Govaert (Université de Technologie de Compiègne) and **Mohamed Nadif** (Université de Metz Ile du Saulcy)

10:15 - 10:45 Double Conjugated Clustering Applied to Leukemia Microarray Data, **Stanislav Busygin**, Gerrit Jacobsen, Ewald Kraemer (Contentsoft AG, Munchen)

10:45 - 11:00 Break

Session II: Information-Theoretic Clustering

11:00 - 11:30 Entropy Based Clustering for High Dimensional Genomic Data Sets, **Donglin Liu** and Gautam Singh (Oakland University)

11:30 - 12:00 An Information-Theoretical Approach to Clustering Categorical Databases Using Genetic Algorithms, Dana Cristofor and **Dan A. Simovici** (University of Massachusetts at Boston)

12:00 - 12:30 Cluster Initialization and Clusterability Detection, Scott Epter, Mukkai Krishnamoorthy, and **Mohammed Zaki** (RPI)

12:30 - 1:45 Lunch

Session III: Clustering Large and High-Dimensional Data

1:45 - 2:15 Using Low-Memory Approximations to Cluster Very Large Data Sets, **David Littau** and Daniel Boley (University of Minnesota)

2:15 - 2:45 Refining Clusters in High Dimensional Text Data, **Inderjit S. Dhillon**, Yuqiang Guan (University of Texas), and J. Kogan (University of Maryland)

2:45 - 3:15 Comparison of Agglomerative and Partitional Document Clustering Algorithms, **Ying Zhao** and George Karypis (University of Minnesota)

3:15 - 3:30 Break

Session IV: Nearest Neighbor and Geometric Techniques

3:30 - 4:00 Making the Nearest Neighbor Meaningful, **Daniel Tunkelang** (Endeca)

4:00 - 4:30 A New Shared Nearest Neighbor Clustering Algorithm and its Applications, Michael Steinbach, Vipin Kumar and **Levent Ertoz** (University of Minnesota)

4:30 - 5:00 How to Partition a Low-Dimensional Data Set Into Subsets of Different Geometric Structures, **Gilad Lerman** (Courant Institute)

Papers