D I M A C S

Center for Discrete Mathematics and Theoretical Computer Science

DIMACS SPECIAL YEAR ON MASSIVE DATA SETS: 1997 ANNUAL REPORT

DIMACS Center • Rutgers University • 96 Frelinghuysen Road • Piscataway, NJ 08854-8018 • USA
TEL: 732-445-5928 • FAX: 732-445-5932 • EMAIL: `center@dimacs.rutgers.edu`
WWW: `http://dimacs.rutgers.edu/` • TELNET: `telnet info.rutgers.edu 90`

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs - Research, Bell Labs, NEC Research, and Telcordia Technologies (formerly Bellcore)

Overview:

The DIMACS special year on massive data sets began with prespecial year activities in 1996-97 and is continuing in 1997-98 as a full-fledged special year.

Organizers:

- Robert Sedgewick, Princeton University (chair)
- Jon Bentley, Bell Labs
- Martin Farach, Rutgers University
- Joan Feigenbaum, AT&T Labs - Research
- Laney Kulsrud, IDA-CCR, Princeton

Sponsors:

The National Science Foundation
The New Jersey Commission on Science and Technology
AT&T Labs - Research
Bell Labs
Bellcore
Rutgers University
Princeton University

Theme of Program:

The emergence of the computer as an essential tool in scientific research and an essential ingredient in commercial systems has led to the emergence of massive amounts of data: Data from astrophysical models, genetic sequencing, geographic information systems, world-wide ecological monitoring, weather prediction, telecommunications applications, biological models for medical applications, government information systems. Researchers in all of these applied areas currently face daunting computational problems in organizing and extracting useful information from these massive data sets. DIMACS' special year is seeking to stimulate new research to address these problems. We are seeking to learn common properties among diverse applications, identify fundamental computational problems that transcend particular applications, develop and analyze new practical algorithms, and establish partnerships with researchers in other areas, with the goal of developing solutions for specific scientific and commercial problems.

Workshops, Miniworkshops and Special Events, 1997-98

- **Session on Mathematical, Statistical, and Algorithmic Problems of Very Large Data Sets**
  - Dates: January 8, 1997
  - Location: American Mathematical Society Meeting, San Diego, CA
  - Organizers: Joan Feigenbaum, AT&T Labs - Research; Fred Roberts, Rutgers University

  In an effort to acquaint the mathematical sciences community with the problems and challenges in this area, the Division of Mathematical Sciences and DIMACS organized a session on massive data sets at the San Diego meeting of the American Mathematical Society. This session described some of the challenging mathematical, statistical, and algorithmic problems inherent in organizing and using enormous amounts of data. The emphasis was on basic issues that transcend particular applications. Speakers explained why existing mathematical, statistical, and algorithmic methods break down on the enormous data sets that scientists and technologists now encounter regularly and attempted to delineate the boundaries at which these breakdowns occur. The session included a discussion of specific DIMACS and DMS programs that focus on huge data sets.

The speakers were: Laney Kulsrud, Institute for Defense Analyses, Center for Communications Research, Princeton; David Karger, Massachusetts Institute of Technology; Sarbani Basu, Aarhus University; and Ken Church, AT&T Labs - Research.

- **Miniworkshop on Exploring Large Data Sets Using Classification, Consensus, and Pattern Recognition Techniques**
  - Dates: May 29-30, 1997
  - Location: Rutgers University
  - Attendance: 58
  - Organizers: F. R. (Buck) McMorris, University of Louisville, and Ilya Muchnik, Rutgers University

  This two day workshop featured the presentation of 21 papers from three separate research areas: academic research in methods, academic research in applications, and industrial research.

  The papers on methods covered the latest results in DM/TCS approaches (convex Boolean functions, hierarchical bases, support vectors, enhanced search, etc.) and the applied papers related to challenging high-tech issues (QSAR, computational finance, very large networks, etc.).

  ¿From the discussions at the workshop it appears that there is not much experience in systematic C2P-supported handling of massive data sets. The large data sets currently maintained by even an advanced company such as AT&T amount to a dozen Gb with just elementary statistical tools implemented at present. The C2P theoretical support under development appears oriented mostly to thousands, not millions of entities. To achieve linear or sub-linear complexity of the algorithms, the theories and methods seem to call for a two-level organization. On the first level, clusters of relatively small samples should be found and analyzed. Next the cluster characteristics should then be extended onto the entire data set.

  Additional topics discussed concerned the preprocessing stage of the data set and analyses such as recognition of false data entries, filling in missed data entries, choice of appropriate discretization grid, and consistency among different data sources.

- **Workshop on Massive Data Sets in Telecommunications**
  - Date: October 13-15, 1997
  - Location: Rutgers University
  - Attendance: 70
  - Organizer: Joan Feigenbaum, AT&T Labs - Research

  Due to the enormous amounts of data involved in telecommunications systems, the telecommunications industry poses some of the most challenging massive data sets problems for researchers in the mathematical sciences. Many classical algorithmic and mathematical problems are relevant, including clustering, compression, error-detection and -correction, nearest-neighbor searching, and shortest-path and connectivity computations.

This workshop attracted many participants from Bell Labs, AT&T Labs and Bellcore as well as a contingent from academia. Dealing specifically with the problems of the telecommunications industry, it was focused towards such industry specific problems as network operations, management, billing, fraud detection, quality control, marketing, information retrieval, and law enforcement.

Events Planned for the Rest of 1997-98

- **Workshop on Mathematical Methods for Massive Data**
  - Dates: April 27-28, 1998
  - Location: Princeton University
  - Organizers: Helene Kulsrud, Institute for Defense Analyses; Robert Grossman, University of Illinois
- **Workshop on Astrophysics and Algorithms: Massive Astronomical Data Sets**
  - Dates: May 6-8, 1998
  - Location: Princeton University
  - Organizers: Michael Vogeley, Princeton University; David Spergel, Princeton University; Don Sun, Bell Labs; Tony Tyson, Bell Labs
- **Workshop on External Memory Algorithms and Visualization**
  - Dates: May 20-22, 1998
  - Location: Rutgers University
  - Organizers: James Abello, AT&T Labs; A.Birilis, AT&T Labs; Steve Mahaney, NSF; Yossi Matias, Bell Labs; T.Munzner, Stanford; W. Sweldens, Bell Labs; Jeff Vitter, Duke University; Kiem-Phong Vo, AT&T Labs.
- **9th Annual Conference on Combinatorial Pattern Matching**
  - Dates: July 20-22, 1998
  - Location: Rutgers University
  - Organizers: Martin Farach, Rutgers University; Udi Manber, University of Arizona

Events Under Development

- **Workshop on Dimensionality, tentative March 1998**
- **Workshop on Biological Databases, tentative Spring 1998**
- **Implementation Challenge, tentative April 1998**
- **Workshop on Biological Applications, tentative June 1998**
- **Workshop on Analysis of Algorithms, tentative July 1998**

Selected Research Results

Optimization Algorithms for Separable Functions with Tree-like Adjacency Variables and their Application to the Analysis of Massive Data Sets Ilya Muchnik (V) and Vadim Mottl (V) developed a new method for very large image, signal and sequence analysis. The base of the method is an extended Bellman function structure which can be efficiently optimized on trees. Many problems from image analysis involve massive data sets. Such a problem can be reformulated as a mapping problem - to transform an original image with a lot of a shape details into "a simple" shape image which is a picture filtered from the original one (in a smoothing problem) or a "symbolic map" which represents the image by a planar graph of a small number of connected fragments-regions each of which is labeled by "a pattern shape"(in a segmentation problem). Such a mapping problem is naturally formulated by using an objective function which depends on a huge number of variables associated with nodes and edges of the basic grid on which the original image is presented. A similar approach is useful for signal (1-dimensional image) analysis and for sequence analysis. The characterization of the above application problems as problems of global optimization is useful in helping us understand them, but on the surface solving these seems unrealistic because of their complexity. However, although the number of variables is large, they are organized in the geometrical lattice-graph with a very simple structure of neighborhoods for both node and edge variables. This allows decomposition into a sequence of optimization subproblems on trees. The crucial idea of the approach is that the optimization problem on a tree is solvable by a dynamic programming approach. The procedure is based on a recursive representation of the corresponding extended Bellman function. The results of Muchnik and Mottl have been applied to real data and the ideas have been extended to apply to sequence alignment problems.

Combinatorial Optimization in Clustering To find some part of a data set which is stable, has a particular property and can be separated from the complementary part of a very heterogeneous large data set, is a core of the apparatus of analysis which is called clustering. There are two main groups of clustering models - statistical and combinatorial. The statistical group is the most popular in data mining today. At the same time, combinatorial models have some value in analysis of massive data sets, because they focus on designing efficient procedures; their results can be interpreted more naturally because they have a simple logical characterization; the results have an extreme characteristic which achieves "the global best value"; and they work well without any assumption of a potential generator of the data. Y. Kempner, Boris Mirkin (V), and Ilya Muchnik (V) have focused on the simplest combinatorial clustering models oriented to massive data set applications. The base of several of them are seriation procedures. They have characterized a new type of set-function (which they call quasi-convex) for which seriations become both an efficient apparatus to find optimal clusters and a good illustration of their structure. Interestingly, the seriations give a dual characterization of the data and bring additional insight into the uncovered structure of the data.

Procedures for On-Line Detection of Seismic Signals Massive data set problems arise when we monitor seismic events and attempt to determine if they are natural earthquakes or underground nuclear tests. Similar problems arise in industrial seismic data analysis related to discovering oil and estimating the amount discovered. In an attempt to establish partnerships with seismologists, DIMACS invited Oleg Kedrov, Director of the Explosion Seismology Laboratory at the Institute of Seismology in the Joint Institute of Earth Physics in the Russian Academy of Sciences, for a month-long visit. Kedrov obtained results on the automatic detection of seismic signals at three-component stations and arrays. In collaboration with L.A. Polikarpova & G.M. Steblov, Kedrov described a new algorithm of on-line signals' detection at a 3-C station based on frequency-time (f-t) and polarization analysis of seismic oscillations. Analysis of detections in the f-t domain using polarization type of the waves and rectilinearity level of the oscillations allows one to reject about 95% of false detections.

A Greedy Algorithm for Logical Data Analysis The Logical Analysis of Data (LAD) is a relatively new methodology aimed at detecting structural information about datasets. It was first proposed by Peter Hammer (PM, RU) and developed by the RUTCOR LAD research group. The main characteristic of LAD is the detection of logical patterns which distinguish observations in one class from all the other observations. For a given set of binary data, a *pattern* is a conjunction of literals which covers at least one positive point and covers no negative points. One measure of a pattern is the coverage, i.e., the number of positive points it covers. Because the problem of finding a pattern that maximizes the coverage is NP-hard, one could not expect to find a polynomial time algorithm. Lijie Shi (GS, RU), Peter Hammer (PM, RU) and Alex Kogan (PM, RU) proposed a greedy algorithm for it. The idea is to try to find a literal by conjuncting it to the partial-pattern such that the resulting new partial-pattern will have the greatest number of positive points and simultaneously decrease the number of negative points as much as possible. Shi implemented the algorithm using C++ and determined that the pattern generated by it was always close to optimal.

Pattern Generation in Logical Analysis of Data In the LAD framework (see above), we are given a data set in which some of the data are classified as positive and the rest negative. LAD is a multimoduled algorithm that seeks to ferret out the variables in the data set that are most relevant to determining the data's classifications. LAD operates in three stages. It first binarizes all numerical variables so that all data are converted to 0 or 1. Secondly, LAD processes the binary data to find a set of boolean conjunctions that have certain properties with respect to the data. These conjunctions are called *patterns.* A pattern must output a 1 for at least one (and ideally many) positive data points, and a 0 for most (ideally all) negative data points. Finally, LAD selects a subset of the patterns that collectively outputs a 1 on a sufficiently large number of positive data points. This subset, called a *theory*, is used to classify future data points whose classification is unknown. The success or value of a theory lies in the accuracy of its classification predictions. An implementation of LAD in C++ has been succesfully tested on numerous data sets. Jason Schattman (GS, RU) implemented in C++ an alternative algorithm, proposed by Alex Kogan (PM, RU), for the pattern generation module of LAD. The current LAD implementation computes a complete enumeration of all patterns up to a fixed degree. Generating patterns of high degree in this way is very expensive computationally, for the number of possible patterns is exponential in the number of binary variables. The goal of this work was to implement a pattern finding algorithm that would find high-degree, high-quality patterns in a computationally efficient way. The algorithm builds patterns one conjunct at a time, adding conjuncts until the conjunction covers sufficiently few negative points. While the pattern is being built, the next conjunct is chosen from the set of candidate conjuncts using one of several heuristics. This approach to pattern generation made possible the identification of very high degree patterns, which could not have been found using current LAD in reasonable time. On most data sets, the algorithm required no more than a few seconds to find enough patterns to cover the entire positive data set. Most importantly, Kogan and Schattman found that many of the patterns found also had very high coverage of the positive data points, which translated into high prediction accuracy on future unclassified data.

Time-series Similarity Problems and Well-separated Geometric Sets. In data mining applications, one frequently needs to determine how similar two time series are. Bela Bollobas (V), G. Das, D. Gunopulos, and H. Mannila analyzed a model of time-series similarity that allows outliers, different scaling functions, and variable sampling rates. They obtained several deterministic and randomized algorithms for computing this notion of similarity. The algorithms are based on tools from computational geometry.