Field: Mathematics/Applied Mathematics/Computer Science

Planning Group Members:
Hiroki Arimura, Graduate School of Information Science and Technology, Hokkaido University
Jörg Rothe, Institute for Computer Science, University of Duesseldorf

Session Topic:
Computational Challenges of Massive Data Sets

Some of the greatest challenges we face in computer science are caused by the rapid growth of data sets. How can one deal with huge amounts of data, find or access useful information, process the data, or compress it such that no substantial information is lost? These are some of the urgent problems at the foremost frontier of research in various fields of computer science, including bioinformatics, image analysis and signal processing, statistical modeling, data mining, data compression, computational complexity, coding theory, and machine learning. This session presents recent advances in some of these fields that each are concerned with the computational challenges arising from massive data sets.

Multimedia Data Processing (Bernhard Grill): Data sets of huge sizes impose severe management, search, and retrieval challenges. Thus, algorithms and tools are needed for identifying and classifying multimedia contents automatically and efficiently, such as tools for handling audio and video signals. Other problems concern the low bitrate coding of audiovisual data in realtime. Using more efficient compression algorithms, multimedia data can be transmitted over very small channels, thus increasing the transmission capacity of the communication network used. This is particularly important when the transmission frequency bandwidth is limited, e.g., in wireless communication. Processing high-definition video material with huge uncompressed data rates poses another great challenge.

Data Assimilation for Earth Sciences (Tomoyuki Higuchi): Numerical simulation is one of the primary means to study complex phenomena in earth sciences. The inherent uncertainty of large and complex dynamic systems, due to the stochastic nature of dependent variables in the systems, requires new methods to analyze the behavior of such systems. Data assimilation aims at combining a large-scale numerical simulation of dynamic systems with a massive data set of observations obtained from large-scale satellite-based or ground-based measurement systems with the aid of statistical modeling techniques. This talk presents a new approach to data assimilation that handles non-linear and non-Gaussian disturbance behaviors in order to realize quantitative predictions of real world problems.

Machine Learning for Brain-Computer Interfacing (Klaus-Robert Müller): Brain-computer interfacing makes use of brain signals for various purposes, such as the control of objects, spelling, etc. This talk surveys brain-computer interfacing from a machine-learning and signal-processing perspective. Here, we are faced with data sets whose size is a truly enormous challenge: Multivariate, strongly noise-contaminated data streams are to be processed in realtime, and neuroelectrical activities are to be accurately differentiated. This talk presents the Berlin Brain-Computer Interface, which is based on EEG signals, and takes the audience all the way from the measured signal to the application.

Data Mining from Massive Data Sets (Einoshin Suzuki): Data mining studies efficient methods for finding patterns or rules in huge data sets. Traditionally, data mining mainly focused on the discovery of "common rules" that frequently appear in a given data set. Exception discovery, a recently emerged subfield of data mining, takes a different approach. It aims at finding exceptions or anomalies, which may lead to unexpected consequences from normal preconditions. This was useful in various applications, e.g., in network intrusion detection or automatic diagnosis from medical data. This talk presents recent developments in exception discovery, including prediction and visualization of chronic hepatitis data, and mining spatio-temporal patterns from multiple video streams of soccer games.