technische universität
dortmund

fakultät für
informatik

Dortmund, 1. Juli 2011

EINLADUNG

An der Fakultät für Informatik wird

Kanishka Bhaduri und Kamalika Das
NASA Ames Research Center

einen Vortrag halten über

Scaling Data Mining Algorithms; Large Scale Earth Science Data

ORT:	E23, OH14, Hörsaal
ZEIT:	Donnerstag, 14. Juli 2011, 16:15 Uhr

ZUSAMMENFASSUNG:

Strategies for Scaling Data Mining Algorithms

In today's world, data is collected/generated at an normous rate in a variety of disciplines starting from mechanical systems e.g. airplanes, cars, etc., sensor networks, Earth sciences, to social networks e.g. facebook. Many of the existing data analysis algorithms do not scale to such large datasets. In this talk, first I will discuss a technique for speeding up such algorithms by distributing the workload among the nodes of a cluster of computers or a multicore computer. Then, I will present a highly scalable distributed regression algorithm relying on the above technique which adapts to changes in the data and converges to the correct result. If time permits, I also plan to discuss a scalable outlier detection algorithm which is at least an order of magnitude faster than the existing methods. All of the algorithms that I discuss will offer provable correctness guarantees compared to a centralized execution of the same algorithm.

Regression Algorithms for Large Scale Earth Science Data

There has been a tremendous increase in the volume of Earth Science data over the last decade. Data is collected from modern satellites, in-situ sensors and different climate models. Information extraction from such rich data sources using advanced data mining and machine learning techniques is a challenging task due to

ZU DIESEN VORTRÄGEN LADEN HERZLICH EIN
DIE DOZENTEN DER FAKULTÄT FÜR INFORMATIK

their massive volume. My research focuses on developing highly scalable machine learning/algorithms, often using distributed computing setups like parallel/cluster computing. In this talk I will discuss regression algorithms for very large data sets from the Earth Science domain. Although simple linear regression techniques are based on decomposable computation primitives, and therefore are easily parallelizable, they fail to capture the non-linear relationships in the training data. In this talk, I will describe Block-GP, a scalable Gaussian Process regression framework for multimodal data, that can be an order of magnitude more scalable than existing state-of-the-art nonlinear regression algorithms.

ZU DIESEN VORTRÄGEN LADEN HERZLICH EIN
DIE DOZENTEN DER FAKULTÄT FÜR INFORMATIK