New grant supports emerging field of massive data analysis and visual analytics

August 6 2008

Enormous amounts of data are being generated in health care, computational biology, homeland security and other areas, but analyzing these massive and unstructured data sets has proven cumbersome and difficult. An emerging research field known as data and visual analytics is helping sift through such mountains of information to find and put together individual pieces of a picture.

The Georgia Institute of Technology has received a five-year grant to lead and coordinate a new initiative that will develop foundational research in massive data analysis and visual analytics. A research team headed by Haesun Park, a professor and associate chair in the Computational Science and Engineering Division of the Georgia Tech College of Computing, will investigate ways to improve the visual analytics of massive data sets through machine learning, numerical algorithms and optimization, computational statistics, and information visualization.

"Developing new and improved mathematical and computational methodologies will further enable systems developers, intelligence analysts, biologists and health care workers to implement new methods to 'detect the expected and discover the unexpected' among massive data sets," Park explained.

The $3 million joint National Science Foundation and Department of Homeland Security grant establishes Georgia Tech as the lead academic research institution for all national Foundations of Data and Visual

Analytics (FODAVA) research efforts. Seven other FODAVA Partnership Awards will be announced later this year, all working in conjunction with eleven Georgia Tech investigators to advance the field.

Over the next five years, the Georgia Tech-led research team will work to establish FODAVA as a distinct research field and build a community of top-quality researchers that will collaborate on research workshops and conferences, industry engagement and technology transfer.

"FODAVA seeks to put an improved science base under one portion of the problem – how can we transform large, complex data sets into reduced computational models or mathematical formalisms that retain the information content while better supporting the human in extracting critical information from the data," said Lawrence Rosenblum, program director for graphics and visualization at the National Science Foundation. "Scientific advances here are critical to future advances in the science of data and visual analytics that will keep us safe and provide technological and commercial advances that benefit mankind."

Georgia Tech's expertise in advanced computer-based analysis, probability and statistics, numerical algorithms and optimization, machine learning, and human-computer interaction techniques provides a strong foundation to lead this new initiative.

Park specializes in using numerical linear algebra and optimization techniques to develop computer-based algorithms that dramatically reduce the dimension and number of data points in massive data sets. Dimension reduction is essential for efficient processing of high-dimension data sets while removing the noise in the data.

Park is especially interested in developing methods for dimension reduction that exploit prior knowledge in the data sets – such as clustered structures and non-negativity. This process is important because it leads

to more accurate classification and prediction results.

Alexander Gray, an assistant professor in the Computational Science and Engineering Division of the College of Computing, has experience developing efficient algorithms that allow statistical and machine learning methods to be applied to massive datasets. He employs ideas from computational geometry and computational physics to statistical computations.

"Reducing the computation time for an analysis from hours to seconds makes all the difference, since data analysis is inherently an iterative and interactive process," explained Gray, also a principal investigator on the project.

Large data sets may also include multiple objects of high dimensionality, such as images, that must be analyzed based on a relatively small number of samples. The mathematical analysis of problems like these requires expertise in statistics and probability methods, which Georgia Tech School of Mathematics professor and principal investigator Vladimir Koltchinskii will contribute to the new initiative.

Once massive amounts of data are collected and processed, relevant information must be pulled from it and presented using visual and interactive means. John Stasko, a principal investigator on this project and professor in the School of Interactive Computing, conducts research in the field of visual analytics.

He heads a team that developed Jigsaw, a visual analytics system that helps analysts better assess, analyze and make sense of large document collections. The system provides multiple coordinated views to show connections between entities extracted from a document collection.

"Jigsaw essentially acts as a visual index of the document collection –

helping analysts identify particular documents to read and examine next," explained Stasko, whose team won the university division of the 2007 Visual Analytics Science and Technology contest using Jigsaw.

Stasko also serves as Georgia Tech's director in the Department of Homeland Security-sponsored SouthEast Regional Visualization and Analytics Center (SRVAC), a regional center created in 2006 to perform research in visual analytics. SRVAC is a partnership between the Georgia Tech and the University of North Carolina Charlotte, and is one of five national university centers connected to the National Visualization and Analytics Center located at Pacific Northwest National Laboratory.

All of the steps involved in massive data analysis and visual analytics – data collection, processing, analysis and visualization – require optimization. Renato Monteiro, a professor in the H. Milton Stewart School of Industrial and Systems Engineering and principal investigator, specializes in this research field.

"This new center provides me the opportunity to apply optimization techniques to new and unique problems and applications that I haven't studied in the past," said Monteiro.

From law enforcement and intelligence gathering to electronic heath records and computational biology, the accurate and timely analysis of massive amounts of information is critical to deeper understanding and effective decision making.

"Collaborations across Georgia Tech's computing, engineering and mathematics disciplines aim to develop better scientific and foundational methods to help practitioners in many different lines of work analyze and interactively explore large data sets more efficiently and effectively," Park added.

Source: Georgia Institute of Technology