Extreme-Scale Visual Analytics

Pak Chung Wong ■ *Pacific Northwest National Lab*

Han-Wei Shen ■ *Ohio State University*

Valerio Pascucci ■ *University of Utah*

The September/October 2004 *CG&A* introduced the term visual analytics (VA) to the computer science literature.[1] In 2005, an international advisory panel with representatives from academia, industry, and government defined VA as "the science of analytical reasoning facilitated by interactive visual interfaces."[2] VA has grown rapidly into a vibrant R&D community offering data analytics and exploration solutions to both scientific and nonscientific problems in diverse domains and platforms. This special issue further examines advances related to extreme-scale VA problems, their analytical and computational challenges, and their real-world applications.

Extreme-scale VA is about applying VA to extreme-scale data. Although the R&D community hasn't yet defined the size of an extreme-scale dataset, much of today's cutting-edge R&D investment in the area is for solving exascale (10^{18}) problems. Among today's most aggressive R&D efforts is the US Department of Energy's (DOE's) Scientific Discovery through Advanced Computing program (SciDAC; http://science.energy.gov/ascr/research/scidac). Two articles in this special issue discuss challenges facing SciDAC scientists.

Size is significant. Analyzing an extreme-scale dataset can rapidly create a set of unsolved and potentially unsolvable problems that challenge foundational VA assumptions regarding algorithms, computation, visualization, user interaction, and databases. The articles in this issue discuss many of these challenges.

The Research Community

The flagship VA conference is the annual IEEE Conference on Visual Analytics Science and Technology (http://visweek.org), which started in 2006. The annual IEEE Symposium on Large-Scale Data Analysis and Visualization (http://ldav.org), now in its second year, emphasizes "algorithms, languages, systems, and hardware that support the analysis and visualization of large data."

Two major R&D communities are investing aggressively in extreme-scale VA. The first is the SciDAC community, which champions solving scientific-discovery problems through exascale data analytics. The second is the e-commerce community, including major online vendors such as Facebook, Google, Yahoo, eBay, and Amazon. This community uses exascale data analytics to tackle its increasingly difficult online data management problems. Much SciDAC R&D employs private clusters called the DOE Leadership Class Facilities. The e-commerce community, on the other hand, focuses on developing extreme-scale data-analytics solutions powered by Apache Hadoop (http://hadoop.apache.org) on public cloud data platforms.

About the Articles

This special issue received submissions from national labs, commercial labs, and universities. After careful reviews by expert reviewers, we selected four articles for publication.

In "A Graph Algebra for Scalable Visual Analytics," Anna Shaverdian and her colleagues argue for a VA graph framework that will facilitate the design of techniques for exascale graph analytics. Their algebraic-based framework aggregates and selects attributes and structural information from very large graphs. These operators aim to find and meaningfully combine information relevant to an analysis. The authors discuss the graph algebra's theoretical foundations. This formalization provides visual analysts a systematic language to document their analysis for reuse and results justification. The authors also present a working

0272-1716/12/$31.00 © 2012 IEEE

implementation over Cytoscape, a popular biological network exploration tool. The implementation demonstrates the algebraic framework and scalable aggregation, using a large social network dataset.

In "Visual Analytics for Finding Critical Structures in Massive Time-Varying Turbulent-Flow Simulations," Kelly Gaither and her colleagues describe how they've employed a dataset of $4{,}096^3$ cells per time slice with 17 time steps, for a total of 1 trillion cells. An automatic feature detection and classification system gleans insight from the data, analyzing the shapes and structures of entropy, an important quality related to flow turbulence. The system identifies coherent structure and important temporal events such as component paths and component birth, death, and reconnection. It also computes statistics such as chord length distribution to provide additional insight into the complex flow fields. The authors implemented the feature analysis on top of VisIt, a popular large-scale visualization system that many application scientists use.

VA has grown rapidly into a vibrant R&D community offering data analytics and exploration solutions to both scientific and nonscientific problems.

In "Geometric Quantification of Features in Large Flow Fields," Wesley Kendall and his colleagues present a VA system for exploring large-scale flow fields. Two scalable components in their back-end processing pipeline support efficient feature queries. One is OSUFlow, a flow-tracing library that can generate millions of field lines in parallel using large-scale supercomputers such as the IBM Blue Gene/P system. The other component is the Scalable Query Interface, which supports parallel querying. Users can use a query tree to query field lines with features related to flow properties such as angle of turn, residence, and other scalar qualities. The authors demonstrate their system's utility using data generated from the Parallel Ocean Program, a high-resolution eddy-resolving ocean circulation model. They also provide a detailed analysis of system performance.

In "Exploratory Visualization Involving Incremental, Approximate Database Queries and Uncertainty," Danyel Fisher and his colleagues stress interactive exploration of very large datasets. With extreme-scale databases, queries can become very slow, making exploratory queries impractical. Researchers can regain interactive speeds by taking progressively larger samples of the dataset, resulting in incremental and approximate visualizations. The authors explore how users interact with the changing incremental visualizations and reveal challenges unique to this visualization mode. They describe guidelines for visualizing changing, approximate results and apply these guidelines by adapting past visualization techniques to convey uncertainty information. Interactive exploration continues to be a top challenge in extreme-scale VA; this research represents one of several approaches to address some of the problems.

Finally, the Visualization Viewpoints department in this issue features an article on the top 10 challenges in extreme-scale VA.[3] In that article, Pak Chung Wong and his colleagues share their views and evaluate the challenges from both technical and social perspectives.

Steve Ashby and his colleagues predicted that all major computational-science components—flops, power, memory, concurrency, storage, I/O bandwidth, and so on—will improve by a factor of 3 to 4,444 by 2018.[4] Beyond hardware architectural advances, the challenges for reaching extreme-scale VA will likely be dominated by visualization, algorithms, databases, and human cognition. As we move beyond the exabyte era (zettabyte data will likely arrive in 2015[5]), these challenges provide opportunities for the VA community to make a positive difference.

Acknowledgments

This work has been supported partly by the US Department of Energy (DOE) Office of Science Advanced Scientific Computing Research under award 59172, program manager Lucy Nowell; DOE award DOE-SC0005036, Battelle Contract 137365; DOE Scientific Discovery through Advanced Computing grant DE-FC02-06ER25770, program manager Lucy Nowell; and US National Science Foundation grant IIS-1017635. Battelle Memorial Institute manages the Pacific Northwest National Laboratory for the DOE under contract DE-AC06-76R1-1830.

References

1. P.C. Wong and J. Thomas, "Visual Analytics," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, 2004, pp. 20–21.
2. J.J. Thomas and K.A. Cook, eds., *Illuminating the*

Path—the Research and Development Agenda for Visual Analytics, IEEE CS, 2005.

3. P.C. Wong et al., “The Top 10 Challenges in Extreme-Scale Visual Analytics,” *IEEE Computer Graphics and Applications*, vol. 32, no. 4, 2012, pp. 63–67.
4. S. Ashby et al., *The Opportunities and Challenges of Exascale Computing: Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee*, US Dept. of Energy Office of Science, 2010; http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf.
5. “Cisco Visual Networking Index: Forecast and Methodology, 2010–2015,” Cisco, 2011; www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.

***Pak Chung Wong** is a chief scientist and project manager at the Pacific Northwest National Laboratory. His research interests include extreme-scale data analytics, visual analytics, social networks, and national security. He’s on the editorial boards of* IEEE Computer Graphics and Applications *and* Information Visualization *and will cochair IEEE VisWeek 2012 and SPIE Visualization and Data Analysis 2013. Wong has a PhD in computer science from the University of New Hampshire. Contact him at pak.wong@pnnl.gov.*

***Han-Wei Shen** is an associate professor in Ohio State University’s Computer Science and Engineering Department. His primary research interests are scientific visualization and computer graphics. He has won the US National Science Foundation’s Career award and the US Department of Energy’s Early Career Principal Investigator Award. He also won the Outstanding Teaching award twice in OSU’s Department of Computer Science and Engineering. Shen has a PhD in computer science from the University of Utah. Contact him at hwshen@cse.ohio-state.edu.*

***Valerio Pascucci** is a professor at the University of Utah’s School of Computing, an associate director at the Scientific Computing and Imaging Institute, and a Laboratory Fellow at the Pacific Northwest National Laboratory. His research interests include scalable algorithms, scientific data analysis, progressive multiresolution techniques in scientific visualization, discrete topology, geometric compression, computer graphics, and applied computational geometry. Pascucci has a PhD in computer science from Purdue University. Contact him at pascucci@acm.org.*