**Large-scale data visualization for data-intensive and high-dimensional scientific data analysis**

Student: Jong Youl Choi (jychoi@cs.indiana.edu)

Advisor: Professor Geoffrey Fox (gcf@indiana.edu)

School of Informatics and Computing

Indiana University, Bloomington, IN 47404

Over the past few decades, the volume of data analysis has been enormously increased in many science domains, including biology, chemistry and sociology, and this data explosion, so-called data deluge, will continue to grow. Discovering useful and meaningful information in large-scale datasets is getting more important in scientific discoveries than ever before. However, such mission is non-trivial and, instead, leaves new challenges and research issues. Not only devising novel data mining algorithms for large-scale data, but also developing efficient parallel and distributed computing systems will play a key role to tackle those problems and expedite scientific discoveries and breakthroughs. In the next decade, we will soon face the era of exascale computing which performs 1018 operations per second. New paradigms and innovative developments of parallel and distributed systems are expected in order to fully utilize the power of exascale computers in many fields of science.

To tackle such challenges, I have conducted my dissertation research mainly in two directions: i) Developing large-scale data visualization algorithms for data-intensive and high-dimensional scientific data analysis. Large-scale visualization is highly valuable in many fields of science and allows us to facilitate scientific discoveries. ii) Implementing distributed and parallel algorithms with efficient use of cutting-edge distributed computing resources, including high-performance clusters, grids, and cloud systems. For this end, I have built a visualization system (Figure 1) which consists of a set of high-performance parallel dimension reduction algorithms, such as traditional Generative Topographic Mapping (EM-GTM), GTM with Deterministic Annealing (DA-GTM), and GTM with interpolation, and a light-weight 3D point visualization client, named PlotViz (Figure 2), by which users can navigate large number of data in a virtual 3D space.



EM-GTM

GTM Interpolation

DA-GTM

High-performance parallel visualization algorithms

PlotViz

**Remote Cluster**

**User Desktop**

 Figure 1. Visualization system architecture Figure 2. 3D point visualization tool, PlotViz

Among many traditional dimension reduction algorithms, I have focused on Generative Topographic Mapping (GTM) for its solid theoretical foundation and superior quality for non-linear data visualization. One drawback, however, is its use of Expectation-Maximization (EM), an optimization method, which can easily be trapped in local optimum, not finding global solutions. I have improved and developed a new variant of GTM, called DA-GTM by applying a novel optimization method called Deterministic Annealing (DA) [[1](#_ENREF_1)]. The new algorithm is more robust against the local optima problem. To the best of my knowledge, this work is the first attempt ever made to apply the Deterministic Annealing (DA) method to GTM. To take further steps to cope with large-scale data problem, I have also improved the GTM algorithm with parallelization [[2](#_ENREF_2)] and interpolation [[3](#_ENREF_3)], respectively. Both of them are designed to maximally utilize a large number of computing nodes concurrently in a distributed cluster system [4].

The visualization system has been successfully applied to real-life data mining projects [[4](#_ENREF_4), [5](#_ENREF_5)], such as a drug discovery project by using cheminfomatics data and a chemogenomic data mining with chemical biology data and the results were presented and published in many high-quality conferences and journals (eg. HPDC, CCGrid, ICCS).

**Thesis Related Publications**

[1] J. Y. Choi, J. Qiu, M. Pierce, and G. Fox, "Generative Topographic Mapping by Deterministic Annealing," presented at the ICCS 2010, Amsterdam, The Netherlands, 2010.

[2] J. Y. Choi, S.-H. Bae, X. Qiu, and G. Fox, "High Performance Dimension Reduction and Visualization for Large High-dimensional Data Analysis," presented at the The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2010), Melbourne, Australia, 2010.

[3] S.-H. Bae, J. Y. Choi, J. Qiu, and G. C. Fox, "Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation," presented at the HPDC'10, Chicago, Illinois USA, 2010.

[4] J. Y. Choi, S.-H. Bae, J. Qiu, G. Fox, B. Chen, and D. Wild, "Browsing Large Scale Cheminformatics Data with Dimension Reduction," in *Proceedings of Emerging Computational Methods for the Life Sciences Workshop (ECMLS) of ACM HPDC 2010 conference*, Chicago, 2010.

[5] J. Y. Choi, S.-H. Bae, J. Qiu, B. Chen, and D. Wild, "Browsing large-scale cheminformatics data with dimension reduction," *Concurrency and Computation: Practice and Experience,* 2011.

**Other publications**

[1] T. Gunarathne, T.-L. Wu, J. Y. Choi, S.-H. Bae, and J. Qiu, "Cloud computing paradigms for pleasingly parallel biomedical applications," Concurrency and Computation: Practice and Experience, 2011.

[2] Y. Yang, J. Y. Choi, K. Choi, M. Pierce, D. Gannon, and S. Kim, "BioVLAB-Microarray: Microarray Data Analysis in Virtual Environment," presented at the IEEE Fourth International Conference on eScience, 2008.

[3] J. Y. Choi, Y. I. Yang, S. Kim, and D. Gannon, "V-lab-protein: Virtual collaborative lab for protein sequence analysis," presented at the IEEE Workshop on High-Throughput Data Analysis for Proteomics and Genomics, Workshop at BIBM 2007, 2007.