Tools for Large-Scale Spatial Simulation Design and Analysis

Johannes Gehrke
Department of Computer Science
Cornell University

We are witnessing an explosion in the generation and availability of data about the world. People are blogging, tweeting status updates, and publishing GPS coordinates of their whereabouts; cameras and sensors freeze the world in space and time by collecting sensory, image, and video data at an unprecedented scale. The challenges in organizing, managing, and mining this data are considerable. In my talk, I will survey two classes of challenges: Challenges that are of algorithmic nature and challenges in usability and programming.

In the first part of my talk I will discuss the challenges of building a useful platform for data-intensive tasks, focusing on the challenges of transportation networks. Simulated virtual environments are already a major enabling technology for engineering research. These environments allow engineers to simulate intelligent behavior, which is crucial for understanding complex systems such as transportation networks, ecosystems, and economies. The economic and environmental impact can be huge. For example, traffic congestion cost $87.2 billion in the United States in 2007, wasting 2.8 billion gallons of fuel and 4.2 billion hours of time. Traffic simulations may help us avoid congestion, consequently reducing air pollution and improving human health. Tomorrow's smart cities will be "made safe, secure environmentally green, and efficient because all structures -- whether for power, water, transportation, etc. are designed, constructed, and maintained making use of advanced, integrated materials, sensors, electronics, and networks which are interfaced with computerized systems comprised of databases, tracking, and decision-making algorithms." []

However, at the heart of all these structures lie design decisions --- what is the impact of various new technologies, how are lifestyle changes of the population impacting our environment, what is the impact of a new transportation option on traffic patterns. This plethora of design decisions requires simulation environments of unprecedented flexibility, where the environment can be programmed in a high-level enough language to introduce a multitude of various actors into the simulation without requiring any low-level optimization or infrastructure extension..

Unfortunately, recurring technical challenges make it difficult to fully exploit the power of these simulations. In particular, it is hard to flexibly program the simulation logic while scaling up the number entities in the simulation in so-called behavioral simulations that are at the core of simulating the city of the future. Current systems either offer high-level programming abstractions, but are not scalable, or achieve scalability by hand-coding particular simulation models using low-level parallel frameworks, such as MPI. While there has been a great deal of work in scaling up general scientific simulations – particularly those defined by systems of partial differential equations – it is difficult to adapt these optimizations to behavioral simulations without the involvement of skilled high-performance computing experts. Additionally, our increasing reliance on cloud platforms is creating new opportunities to run these simulations cheaply, but also new challenges due to the shared environment and the reduced reliability of individual components of the computational infrastructure.

We have been working on a simulation platform that allows engineers to create large-scale behavioral simulations rapidly without specialized computer science training. Our platform balances high-level programmability and high performance. Just as databases have given organizations new tools for exploring data, our simulation platform enables engineers to rapidly prototype and explore complex behaviors [3]. I will discuss the motivation of our platform, its programming model, and will give some example applications.

In the second part of my talk, I will discuss the algorithmic challenges that arise if we were to equip every driver with a camera and would suddenly be flooded by continuous streams of images. I will explore this problem through the lens of high-dimensional similarity search, a specific data mining primitive. Given a set of high-dimensional binary vectors, similarity search finds all pairs of vectors whose similarity exceeds a user-defined threshold. This step is a key component in applications such as such as image clustering, near duplicate documents detection, 3D scene reconstruction, similar music and video retrieval, community mining, and personalized recommendations. The class of techniques that I will discuss is based on computing sketches of objects. These sketches use very little space, but have the ability that with high probability they approximate pairwise distances very well. I will introduce different types of sketches and discuss how they can be applied to similarity search [1,4].

[1] Alexandr Andoni and Piotr Indyk. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. Communications of the ACM, vol. 51, no. 1, 2008, pp. 117-122.

[2] Robert Hall. The Vision of a Smart City. 2nd International Life Extension Technology Workshop, Paris, France, September 2000.

[3] Guozhang Wang, Marcos Antonio Vaz Salles, Benjamin Sowell, Xun Wang, Tuan Cao, Alan J. Demers, Johannes Gehrke, Walker M. White: Behavioral Simulations in MapReduce. PVLDB 3(1): 952-963 (2010)

[4] Jiaqi Zhai, Yin Lou, Johannes Gehrke: ATLAS: A Probabilistic Algorithm for High Dimensional Similarity Search. In Proceedings of the 2011 ACM SIGMOD Conference. Athens, Greece, June 2011.