Algorithmic Design Considerations for Geospatial and/or Temporal Big Data

Terence van Zyl

Abstract

In order to frame the geospatial temporal big data conversation, it is important to discuss them within the context of the three Vs (velocity, variety, and volume) of big data. Each of the Vs brings its own technical requirements to the algorithmic design process, and each of these requirements needs to be considered. It is also important to acknowledge that some of the challenges facing the broader big data community have always existed within the geospatial temporal data analytics community and will always continue to do so. Especially relevant are those big data challenges relating to data volume as presented by large quantities of either raster data, point clouds, and even vector data. Spatial data mining has long endeavored to unlock information from large databases with spatial attributes, and in these cases, algorithmic approaches have been adapted to overcome the data volume. Although the problem of big data is one that is well acknowledged and long studied, it is worth gaining a deeper insight and a more formal and rigorous treatment of the subject as is presented by the opportunity of a sudden awareness of spatial big data by the broader data community. Spatial data can be categorized into three major forms: these being raster, vector, and areal. Historically, it has been the case that raster data presented itself as a large volume challenge. It is clear that this historical trend is changing, and none of these categories maps neatly to any of the big data's Vs. For example, a large volume of vector data is now plausible if the Internet of Things is considered, and these data could also place velocity constraints on the algorithms if near real-time processing is required. Additionally, high-variety unstructured data may arrive at high velocity or any other of the many permutations. What is clear across all these permutations of the big data Vs is that considerable consideration needs to be given to the time and space complexity of the algorithms that are required to process these data. In addition, each of the three Vs places added constraints on the others, and increasingly, the three Vs need to be considered together. For example, unstructured data increases the time complexity of algorithms needed to process the data chunks, while, for instance, high volumes of the same unstructured data increase space complexity. To gain a true sense of the overall challenge faced by the geospatial big data community, couple these classical challenges of big data with the added time and space complexity of spatial data algorithms. First, it is important to note that the independent identical distribution (IID) is not a reasonable assumption for either temporal or spatial data. The reason for this assumption failing is that both of these cases consider data that is auto-correlated. In fact, the first rule of geography is this fact exactly. As a result of not being able to make an IID assumption in

most cases, the time complexity of spatial and temporal algorithms is higher than their traditional counterparts. For example, Spatial Auto-Regression is more complex than Linear Regression, Geographically Weighted Regression is more computationally demanding than Regression, and Co-Location Pattern Mining requiring spatial predicates is more complex than Association Rule Mining. In addition, ignoring the spatiotemporal autocorrelation in the data can lead to spurious results, for instance, the salt and pepper effect when clustering. The solution to the big data challenge is simple to describe yet in most cases is not easily tractable. Simply put it is important insofar as it is possible to minimize space complexity aiming for at most linear space complexity and target a time complexity that is log linear if not less. However, this is often not possible and other techniques are required. All is not lost and spatial data does not only present increased challenges in the big data arena but also provides additional exploitable opportunities in overcoming some of the big data challenges. For example, spatial autocorrelation allows for aggregations and filtering of data within fixed windows so as to reduce the total number of points required for consideration without excessive loss of information. It also allows the algorithm designer to consider points at sufficient distance as a single cluster thus reducing the number of computations.