Where and How should Streaming Sensor Data be Processed?

Anand Ranganathan
IBM TJ Watson Research Center
Hawthorne, NY 10532, USA
arangana@us.ibm.com

Sensors today produce huge amounts of real-time data. This data needs to be processed in real-time as well so as to generate useful knowledge and insights for the purpose of making rapid decisions. There are different places where this data can be processed. This includes on the sensor itself, within a sensor network, at a gateway with which the sensor network communicates, and in a centralized data processing center. There are a number of factors that influence the decision of where different kinds of processing should be performed on different kinds of data. Some of these factors include the volume and rate of data produced by the sensors, the data processing latency requirements, the complexity of the processing, security and privacy considerations, administrative or organizational boundaries, whether there is a need for accessing data from databases and other relatively static data sources, etc. In this paper, we explore some of these issues in the context of processing sensor data for purposes of traffic monitoring and prediction.

The problem of where to perform different kinds of processing on sensor data is particularly apparent in the domain of Intelligent Transportation Systems (or ITS). Intelligent Transportation Systems is an umbrella term encompassing sensor, communications and computing technologies to manage existing infrastructure and transportation systems more efficiently, and hence contribute to the reduction of congestion. An important development within ITS has been the emergence and installation of different kinds of sensor technologies for collecting data on the state of the transport system [1]. These include both floating car data (FCD) as well as fixed sensor data. FCD represents the location of vehicles collected by mobile sources, such as GPS devices installed in vehicles or cellular phones. Fixed sensor data includes data from video cameras, loop detectors, toll booths, etc. These different sensors have great potential to provide the large amounts of data that is needed to support real time management of traffic systems.

There are several steps of processing that are required to go from raw sensor data to estimates of current traffic on different road links, as well as predictions of future traffic conditions. Let us consider the case of using GPS data from cellphones and vehicles. The basic steps in processing this data include a) cleaning and de-noising the data, e.g., to filter out cellphones that are not in moving vehicles, b) map-matching of the GPS location to the road network, c) calculating the average speed of individual vehicles, and d) aggregating the data to produce traffic statistics (speed, flow, density) on different links.

These different data processing steps can occur in different places. For example, steps (a) and (b) could potentially be performed on the GPS device itself, if the device had access to the complete map information of the city. This is possible in the case of GPS devices built into vehicles, but may not be possible for cellphones. Steps (c) and (d) could be performed at a centralized data processing facility. Such a facility may make use of a stream processing infrastructure (e.g. IBM InfoSphere Streams [3]). Such an infrastructure can handle high-complexity analyses like traffic prediction, which may make use of traffic models built using machine learning techniques, or using simulation approaches.

Apart from processing performed on the device and in a centralized location, some processing can also occur at the gateway or mobile switching center, especially in the case of cellphone based GPS data. This step may be performed by the cellphone service provider or an authorized party. The processing performed here may include aggregation and anonymization of the raw GPS data for the purpose of preserving the privacy of individual cell phone users.

Another approach to split the data processing is to perform all the processing on a centralized infrastructure. In this case, the sensors are just responsible for sending the raw data. In earlier work [2], we had described a centralized, stream-processing approach for doing all the processing of real-time vehicular GPS data. We had used the IBM InfoSphere Streams product [3], which is a new product that supports high performance stream processing. It offers both language and runtime support for improving the performance of streaming applications via a combination of optimized code generation, pipelining and parallelization. It also supports a component-based programming model that simplifies the development of complex applications. The advantage of using such a centralized approach is that it reduces the computational requirements on the sensors. In addition, since all the data processing happens in one place, it is easier to adapt or extend the processing to meet changing requirements.

In this particular example, there was no sensor network per se. However, if data was obtained from a sensor network (e.g. environmental data on the condition of roads), this network would be another possible place to do data processing. Such in-network data processing can help reduce communication and/or computational needs of the sensor network.

In summary, there are different places where sensor data can be processed to produce useful information. So far, however, the partitioning of the processing has tended to be dictated by organizational or administrative boundaries. More research needs to be done to determine what is the optimal way of splitting the processing to meet performance, privacy, functionality and other requirements.

References

[1] C. Antoniou, R. Balakrishna, and H. Koutsopoulos. Emerging data collection technologies and their impact on traffic management applications. *ASCE Journal of Transportation Engineering*, 2009.

[2] A. Biem, E. Bouillet, H. Feng, H. Koutsopoulos, C. Moran, A. Ranganathan, A. Riabov, and O. Verscheure. IBM InfoSphere Streams for Scalable, Real-Time, Intelligent Transportation Services. In *SIGMOD '10 : Proceedings of the ACM International Conference on Management of Data*, 2010.

[3] IBM Infosphere Streams. http://www-01.ibm.com/software/data/infosphere/streams/.