Qualification and Summarization of Environmental Sensors Data Streams

Mots clés :

- Directeur de thèse : amara AMARA
- Co-encadrant(s) :
- Unité de recherche : Laboratoire d'Informatique Signal et Image Electronique et Télécommunication
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

An environmental monitoring process consists of regular observation and information collection about the environment, and regular recording and analysis of the observed data. It involves returning the inferred information and conclusions to the environment explorer. The returned information will be used by the network explorer for making decision. This process is accomplished by an environmental monitoring system. In order to collect data, the first step is to place a set of specific sensors in the environment. These sensors are used to control and supervise the environment. They continuously measure the environmental information, like the temperature, humidity, pressure, etc. The collected data should be then exploited and interpreted, in order to come out with the useful and reliable information. This represents the data analysis process, which helps the environment explorer to make right decisions. After the data collection and analysis phases, the monitoring system communicates the inferred information to the network explorer who makes decisions based on these deductions. This whole process, from the data collection to the data analysis, will lead to two key problems: the amount of the data and the data quality. In fact, a sensor generates the data in the form of a stream, which consists of a huge volume of data sent to the monitoring system in a continuous way. The arrival rate of the data is very high compared to the available treatment and storage capacities. The monitoring system is thus facing a significant amount of data for which a permanent and exhaustive storage is very expensive and sometimes impossible. That is why we need to treat the data stream in a single pass, without storing it. However, for a particular stream, it is not always possible to predict in advance all of the processing to be performed. Due to this problem, new challenges related to data mining and analysis have appeared. To overcome this problem, it is necessary to save some of this data in a compact structure called “summary”. There are a variety of techniques that can be used to build a data stream summary, among them: the histograms, wavelets, sketch and sampling algorithms. The challenge is to decide what to store in this summary, and how to ensure that the summary satisfies the requirements of the application, while respecting the system resources available. Consequently, our first challenge is how to build a data stream summary with a good performance. On the other hand, in a real world such as the sensors environments, the data are often dirty, they contain noise, erroneous or missing values. This is due to many factors: local interferences, malicious nodes, network attack, network congestion, limited sensor precision, harsh environment, sensor breakdown, sensor malfunction, miscalibrating and insufficient battery power of the sensor. And, as in the data analysis process the conclusions and decisions are based on the data, this leads to defective and faulty results if the data are not clean. One solution to overcome this problem is to use sensors with high precision to could assume that the arising errors are small, and to deploy redundant sensors to cover the breakdown of a sensor. However, this approach is very expensive as it requires very high costs for the sensors. Another approach will be presented in this thesis. It consists of using some methods and algorithms to first evaluate the data quality, and then, improve this quality to be able to obtain reliable and efficient results. Our work is supported by the FUI 17 project WAVES. The goal of this project is to design and develop a platform for distributed management of massive structured and unstructured data streams. The considered use case concerns the real-time supervision of the water distribution network.