



Tutoriels

1. **Georges Hébrail** , Professeur à Telecom-ParisTech

Title: An introduction to data stream querying and mining

Abstract

Human activity is nowadays massively supported by computerized systems. These systems handle data to achieve their operational goals and it is often of great interest to query and mine such data with a different goal: the supervision of the system. The supervision process is often difficult (or impossible) to run because the amount of data to analyze is too large to be stored in a database before being processed, due in particular to its historical dimension.

This problem has been studied intensively for several years, mainly by researchers from the database field. A new model of data management has been defined to handle “data streams” which are infinite sequences of structured records arriving continuously in real time. This model is supported by newly designed data processing systems called “Data Stream Management Systems”. These systems can connect to one or several stream sources and are able to process “continuous queries” applied both to streams and standard data tables. These queries are qualified as continuous because they stay active for a long time while streaming data are transient. The key feature of these systems is that data produced by streams are not stored permanently but processed ‘on the fly’. Note that this is the opposite of standard database systems where data are permanent and queries are transient. Such continuous queries are used typically either to produce alarms when some events occur or to build aggregated historical data from raw data produced by input streams.

As data stored in data bases and warehouses are processed by mining algorithms, it is interesting to mine data streams, i.e. to apply data mining algorithms directly to the streams instead of storing them beforehand in a database. This problem has also been studied a lot and new data mining algorithms have been developed to be applicable directly to streams. These new algorithms process data streams ‘on the fly’ but they can also provide results based on a portion of the stream instead of the whole stream already seen. Portions of streams are defined by fixed or sliding windows.

We will provide an introduction to the data stream management and mining field. First, the main applications which motivated these developments will be presented (telecommunications, computer networks, stock market, security, ...) and the new concepts related to data streams will be introduced (structure of a stream, timestamps, time windows, ...). A second part will present the main concepts and architectures related to Data Stream Management Systems. The third part will present the main results about the adaptation of data mining algorithms to the case of streams.

2. **Ludovic Lebart**, Directeur de Recherches CNRS, Telecom-ParisTech

Title: Text Mining

Abstract



3. **Andre Ponce de Leon**, Professeur à l'USP, Etat de São Paulo

Title: Data Mining in Bioinformatics

Abstract

Molecular Biology laboratories have gathered a very large amount of data in sequence and functional genome projects. It is frequently not possible to analyze these data manually. Sophisticated computing techniques are necessary to extract new, meaningful and useful information from these data. Data Mining techniques have been successfully applied in such analysis. Examples of these applications are analysis of gene expression data, recognition of genes in DNA sequences and protein structure prediction. This tutorial will present the main issues on the use of Data Mining techniques in Bioinformatics. The tutorial will start with the introduction of the key aspects of Data Mining, with special emphasis on Machine Learning. Next, the necessary issues of molecular biology for the understanding of the Data Mining applications in Bioinformatics will be described. Usually, biological data needs to be pre-processed before they can be used in a Data Mining process. The main techniques for data pre-processing will be presented. Later, a few applications of data Mining techniques, mainly classification and clustering, to bioinformatics problems will be presented.