Statistics Seminar Series

Data Summarization for Data Mining Applications.

George Kollios
Department of Computer Science
Boston University

Thursday, October 17, 2002, 4:00-5:00pm
Mathematics and Computer Science (MCS) Building, Room 149
111 Cummington Street, Boston

Tea and Cookies at 3:30pm in MCS 153

Abstract: Data summarization and data reduction of very large datasets is an important step in many data mining and data analysis applications. In the first part of the talk I will present methods to approximate multi-dimensional datasets for estimating the selectivity of range queries. These methods can be used for data exploration and database query optimization. The simplest approach to tackle this problem is to assume that the attributes are independent. More accurate estimators try to capture the joint data distribution of the attributes. In databases, such estimators include the construction of multi-dimensional histograms, random sampling, or the wavelet transform. We present a new histogram technique that is designed to approximate the density of multi-dimensional datasets with real attributes. Our technique defines buckets of variable size, and allows the buckets to overlap. The size of the cells is based on the local density of the data. The use of overlapping buckets allows a more compact approximation of the data distribution. We also investigate how to use kernel density estimators on the multi-dimensional query approximation problem.

For directions and maps, please see *http://math.bu.edu/research/statistics/statseminar.html*. For other information, please contact Eric Kolaczyk (*kolaczyk@math.bu.edu*) or the main department office at (617)353-2560.