

RUTGERS UNIVERSITY
DEPARTMENT OF STATISTICS AND BIOSTATISTICS
HILL CENTER #501, BUSCH CAMPUS, PISCATAWAY

www.stat.rutgers.edu

Seminar

Speaker: Aiyou Chen, Bell Lab, Murray Hill, NJ
Title: Distinct counting with a self-learning bitmap
Date: Wednesday February 11, 2009
Time: 3:20 PM
Place: 552 Hill Center

Abstract

Estimating the number of distinct values is a fundamental problem in database that has attracted extensive research over the past two decades, due to its wide applications (especially in the Internet). However, the performance of existing algorithms in terms of relative estimation accuracy usually depends on the unknown cardinalities. In this paper we address the following question: can a distinct-counting algorithm have uniformly reliable performance, i.e. constant relative estimation errors for unknown cardinalities in a wide range, say from tens to millions? Reliability is important in applications. We propose a self-learning bitmap algorithm to answer the question. We demonstrate that with a given memory requirement, the algorithm is not only uniformly reliable but more accurate than state-of-the-art algorithms such as the multi-resolution bitmap (Espan et al) and hyper Loglog (Flajolet et al) algorithms under common practice settings. A rigorous proof of uniformity is ongoing. This is joint work with Jin Cao and Larry Shepp.