Sampling for Big Data: A Tutorial

Graham Cormode
University of Warwick*
g.cormode@warwick.ac.uk

Nick Duffield
Rutgers University / DIMACS†
nick.duffield@rutgers.edu

ABSTRACT

One response to the proliferation of large datasets has been to develop ingenious ways to throw resources at the problem, using massive fault tolerant storage architectures, parallel and graphical computation models such as MapReduce, Pregel and Giraph. However, not all environments can support this scale of resources, and not all queries need an exact response. This motivates the use of sampling to generate summary datasets that support rapid queries, and prolong the useful life of the data in storage. To be effective, sampling must mediate the tensions between resource constraints, data characteristics, and the required query accuracy. The state-of-the-art in sampling goes far beyond simple uniform selection of elements, to maximize the usefulness of the resulting sample. This tutorial reviews progress in sample design for large datasets, including streaming and graph-structured data. Applications are discussed to sampling network traffic and social networks.

1. INTRODUCTION

Research in big data draws from many disciplines, both from areas of analysis (including probability, data analysis, machine learning and algorithms) and also systems (including distributed systems, database architectures, and programming models). This tutorial addresses several target audiences, all of whom may want to benefit from a greater understanding of the use of probabilistic methods for sampling and estimation in big data:

Researchers in big data who background is not primarily concerned with probabilistic methods, but wish to learn about these as a complement to their current expertise.

Researchers with a more classical background in statistics who wish to refocus to apply their expertise to problems in big data.

Researchers in big data methodologies who wish to learn more about current applications of sampling in big data in the field.

2. OUTLINE

Fundamentals of Sampling. Summarization as a means to control resource usage, sampling as special case of summarization. General comparison of sampling with other summarization techniques, including aggregation and sketching: different trade-offs between scalability, flexibility, and ability for post-hoc exploration. Sampling as a mediator between data characteristics. Unbiased estimation and the Horvitz-Thompson approach. Variance and Covariance Estimators. General purpose samples vs. targeted estimators.

Sampling from Streams of Data. Uniform Stream Sampling, and generalizations to weighted streams, via Reservoir Sampling, Sample and Hold and Counting Samples. Inclusion Probability Proportional to Size Sampling, including Threshold Sampling, Priority Sampling, Variance Optimal Sampling, and Bottom-k Sampling Sketch guided sampling, Structure Aware Sampling, Sampling from the distinct objects via ℓ_0 sampling, and ℓ_p sampling.

Hashing and Coordinated Sampling. Hash functions as a source of Permanent Random Numbers (PRN). Hashing algorithms, including Universal Hashing and implementations. Advanced topics may include Minwise hashing, Consistent Weighted Hashing, Hashing with Timeouts, Coordinated Sampling, or Trajectory Sampling

Graph Sampling. Network sampling, ego-net sampling, and why simple edge or node centric sampling fails for more complex properties. Topics may include Biased sampling methods, Random walks, Snowball and respondent driven sampling, forest fire sampling, Graph stream sampling, Horwitz-Thompson estimation in graphs, or Triangle counting.

Sampling Applications. Network Traffic Measurement (Netflow, IPFIX, PSAMP); Sampling and Approximate Database Queries (Priority Sampling in Databases, BlinkDB); Social Networks (Social activity streams, Sampling and information diffusion).

3. TUTORS

Graham Cormode is a Professor in Computer Science at the University of Warwick in the UK.
Nick Duffield is a Research Professor at Rutgers University/DIMACS, New Jersey, USA.

*Supported in part by a Yahoo! Research FREP award.

†Address from September 2014: Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128

KDD'14, August 24–27, 2014, New York, NY, USA.
ACM 978-1-4503-2956-9/14/08.
http://dx.doi.org/10.1145/2623330.2630811.