



Small Summaries for Big Data

Graham Cormode and Ke Yi

Cambridge University Press, 2020

This book is expected to be published in 2020. This book is aimed at both students and practitioners interested in algorithms and data structures for working with very large volumes of data. These techniques are of relevance to people working in big data, data science, and machine learning. The algorithms described have been adopted in systems from tech companies like Google, Apple, Microsoft, Netflix and Twitter, and many more.

This material will be published by Cambridge University Press as *Small Summaries for Big Data* by Graham Cormode and Ke Yi. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. © copyright Graham Cormode and Ke Ye 2019.

Summary

The volume of data generated in modern applications can be massive, overwhelming our abilities to conveniently transmit, store, and index. For many scenarios, it is desirable to instead build a compact summary of a dataset that is vastly smaller. In exchange for some approximation, we obtain flexible and efficient tools that can answer a range of different types of query over the data. This book provides a comprehensive introduction to the topic data summarization, showcasing the algorithms, their behavior, and the mathematical underpinnings of their operation. The coverage starts with simple sums and approximate counts, building to more advanced probabilistic structures such as the Bloom Filter, distinct value summaries, sketches, and quantile summaries. Summaries are described for specific types of data, such as geometric data, graphs, and vectors and matrices. Throughout, examples, pseudocode and applications are given to enhance understanding.

Contents

We are making draft chapters publicly available (this snapshot from October 2019 is the pre-production draft). We welcome comments/corrections by email to the authors.

- Chapter 1 - Introduction
- Chapter 2 - Summaries for Sets
- Chapter 3 - Summaries for Multisets
- Chapter 4 - Summaries for Ordered Data
- Chapter 5 - Geometric Summaries
- Chapter 6 - Vector, Matrix, and Linear Algebraic Summaries
- Chapter 7 - Graph Summaries
- Chapter 8 - Summaries over Distributed Data
- Chapter 9 - Other Uses of Summaries
- Chapter 10 - Lower bounds for Summaries
- References and index