Synopsis Data Structures for Massive Data Sets

Phillip B. Gibbons* Yossi Matias†

1 Introduction

A growing number of applications demand algorithms and data structures that enable the efficient processing of data sets with gigabytes to terabytes to petabytes of data. Such massive data sets necessarily reside on disks or tapes, making even a few accesses of the base data set comparably slow (e.g., a single disk access is often $10,000$ times slower than a single memory access). This short note (see [8] for the full paper) considers data structures for supporting queries to massive data sets, while minimizing or avoiding disk accesses. In particular, we advocate and study the use of small space data structures.

We denote as *synopsis data structures* any data structures that are substantively smaller than their base data sets. A synopsis data structure has the following advantages over a non-synopsis (e.g., linear space) data structure: (a) it may reside in main memory, enabling query responses and data structure updates that avoid disk accesses altogether, (b) it can be transmitted remotely at minimal cost, (c) it has minimal impact on the overall storage costs of a system, (d) it leaves space in the memory for other processing (available main memory is a precious resource for external memory algorithms), and (e) it can serve as a small surrogate for data sets that are currently expensive or impossible to access. Hence a traditional viewpoint in the algorithms literature — that a linear space data structure is a good one — is not appropriate for massive data sets, as such data structures often fail to provide satisfactory application performance.

On the other hand, since synopsis data structures are too small to maintain a full characterization of their base data sets, they must summarize the data set, and the responses they provide to queries will typically be approximate ones. The challenges are to determine (1) what synopsis of the full data set to keep in the limited space in order to maximize the accuracy and confidence of its approximate responses, and (2) how to efficiently compute the synopsis and maintain it in the presence of updates to the data set.

Due to their importance in applications, there are a number of synopsis data structures in the literature and in existing systems. Examples include uniform and biased random samples, various types of histograms, statistical summary information such as frequency moments, data structures resulting from lossy compression of the data set, etc. Often, synopsis data structures are used in a heuristic way, with no formal properties proved on their performance or accuracy, especially under the presence of updates to the data set. Our ongoing work since 1995 seeks to provide a systematic study of synopsis data structures, focusing on performance and accuracy guarantees, even in the presence of data updates.

2 Results

In the full paper, we describe a context for algorithmic work relevant to massive data sets and a framework for evaluating such work. In brief, we combine the PDM external memory model [15] with input/output conventions more typical for the study of (online) data structure problems. Two general scenarios are considered: one where the input resides on the disks of the PDM and one where the input arrives online in the PDM memory. We describe some of our work on synopsis data structures, and highlight results on three problem domains from the database literature: hot list queries, histograms and quantiles, and frequency moments.

Hot list queries. A *hot list* is an ordered set of $\langle$value, count$\rangle$ pairs for the most frequently occurring "values" in a data set. Hot lists are used in a variety of data analysis contexts, including fraud detection, best sellers lists, market basket analysis, selectivity estimation in query optimization, load balancing in parallel query optimization, etc. Note that hot lists are trivial to maintain given sufficient space to hold a full histogram of the data set; however, for many data sets, the histogram would not be a synopsis data structure. We show in [1] that there are no synopsis data structures for estimating even the count of the most

*Information Sciences Research Center, Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974. Email: gibbons@research.bell-labs.com.

†Department of Computer Science, Tel-Aviv University, Tel-Aviv 69978 Israel, and Information Sciences Research Center, Bell Laboratories. Email: matias@math.tau.ac.il. Research supported in part by an Alon Fellowship, by the Israel Science Foundation founded by The Academy of Sciences and Humanities, and by the Israeli Ministry of Science.

frequently occurring value, to within constant factors, over all distributions. On the other hand, we present in [7] new synopsis data structures that are shown both analytically and experimentally to produce more accurate approximate hot lists than previous methods, and perform quite well for the skewed distributions that are of interest in practice. See also [5] for a recent study of a related problem.

Maintaining histograms and quantiles. Histograms approximate a data set by grouping values into "buckets" (subsets) and approximating the distribution of values in the data set based on summary statistics maintained in each bucket (see, e.g., [14]). They are used extensively in commercial databases for selectivity estimation purposes within a query optimizer and in query execution. A common problem with histograms is keeping them up-to-date in the presence of data updates. We present in [10] algorithms for fast dynamic maintenance of highly-accurate approximate histograms for the most widely used classes of histograms. Additional recent work is given in [3, 12].

Frequency moments. Estimating the number of distinct values in a data set is a problem that frequently occurs in database applications. From an algorithmic point of view it demonstrates the advantages of viewing the input online using a synopsis data structure vs. only sampling from the input. Indeed, the algorithms in [6, 1] demonstrate effective synopsis data structures of size $O(\log n)$ for the online problem, whereas $\Omega(n)$ space is required for any sampling-only approach, regardless of the estimator used [3]. Synopsis data structures for other frequency moments are presented in [1], as well as corresponding size lower bounds.

Other results. A recent survey by Barbará *et al.* [2] describes and qualitatively evaluates the state of the art in *data reduction* techniques. Our work on synopsis data structures also includes the use of multifractals and wavelets for synopsis data structures [4, 13], and join synopses for queries on the join of multiple sets. This work is part of the *Approximate query answering (Aqua)* project [9, 11] at Bell Labs; Aqua seeks to provide fast, approximate answers to queries using synopsis data structures.

While synopsis data structures have been proposed and studied for a number of query problems (see the full paper for additional examples), many more open questions remain, and we hope that this short note will motivate others in the algorithms community to study these problems.

References

[1] N. Alon, Y. Matias, and M. Szegedy, *The space complexity of approximating the frequency moments*, in Proc. 28th ACM Symp. on the Theory of Computing, May 1996, pp. 20–29. Full version to appear in JCSS special issue for STOC'96.

[2] D. Barbará et al., *The New Jersey data reduction report*, Bulletin of the Technical Committee on Data Engineering, 20 (1997), pp. 3–45.

[3] S. Chaudhuri, R. Motwani, and V. Narasayya, *Random sampling for histogram construction: How much is enough?*, in Proc. ACM SIGMOD Int'l Conf. on Management of Data, June 1998, pp. 436–447.

[4] C. Faloutsos, Y. Matias, and A. Silberschatz, *Modeling skewed distribution using multifractals and the '80-20' law*, in Proc. 22rd Int'l Conf. on Very Large Data Bases, Sept. 1996, pp. 307–317.

[5] M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. D. Ullman, *Computing iceberg queries efficiently*, in Proc. 24th Int'l Conf. on Very Large Data Bases, Aug. 1998, pp. 299–310.

[6] P. Flajolet and G. N. Martin, *Probabilistic counting algorithms for data base applications*, J. Computer and System Sciences, 31 (1985), pp. 182–209.

[7] P. B. Gibbons and Y. Matias, *New sampling-based summary statistics for improving approximate query answers*, in Proc. ACM SIGMOD Int'l Conf. on Management of Data, June 1998, pp. 331–342.

[8] ———, *Synopsis data structures for massive data sets*, DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, (1998). To appear. Available as Bell Labs tech. rep., Sept. 1998, and at http://www.bell-labs.com/~pbgibbons/.

[9] P. B. Gibbons, Y. Matias, and V. Poosala, *Aqua project white paper*, tech. rep., Bell Laboratories, Murray Hill, New Jersey, Dec. 1997.

[10] ———, *Fast incremental maintenance of approximate histograms*, in Proc. 23rd Int'l Conf. on Very Large Data Bases, Aug. 1997, pp. 466–475.

[11] P. B. Gibbons, V. Poosala, S. Acharya, Y. Bartal, Y. Matias, S. Muthukrishnan, S. Ramaswamy, and T. Suel, *AQUA: System and techniques for approximate query answering*, tech. rep., Bell Laboratories, Murray Hill, New Jersey, Feb. 1998.

[12] G. S. Manku, S. Rajagopalan, and B. G. Lindsley, *Approximate medians and other quantiles in one pass and with limited memory*, in Proc. ACM SIGMOD Int'l Conf. on Management of Data, June 1998, pp. 426–435.

[13] Y. Matias, J. S. Vitter, and M. Wang, *Wavelet-based histograms for selectivity estimation*, in Proc. ACM SIGMOD Int'l Conf. on Management of Data, June 1998, pp. 448–459.

[14] V. Poosala, Y. E. Ioannidis, P. J. Haas, and E. J. Shekita, *Improved histograms for selectivity estimation of range predicates*, in Proc. ACM SIGMOD Int'l Conf. on Management of Data, June 1996, pp. 294–305.

[15] J. S. Vitter and E. A. M. Shriver, *Algorithms for parallel memory I: Two-level memories*, Algorithmica, 12 (1994), pp. 110–147.