Constructing Optimal Wavelet Synopses

Dimitris Sacharidis

Knowledge and Database Systems Lab
School of Electrical and Computer Engineering
National Technical University of Athens
Zographou 157 73, Athens, Greece
dsachar@dblab.ntua.gr

Abstract. The wavelet decomposition is a proven tool for constructing concise synopses of massive data sets and rapid changing data streams, which can be used to obtain fast approximate, with accuracy guarantees, answers. In this work we present a generic formulation for the problem of constructing optimal wavelet synopses under space constraints for various error metrics, both for static and streaming data sets. We explicitly associate existing work and categorize it according to the previous problem formulation and, further, we present our current work and identify its contributions in this context. Various interesting open problems are described and our future work directions are clearly stated.

1 Introduction

Approximate query processing over compact precomputed data synopses has attracted a lot of attention recently as an effective approach for dealing with massive data sets in interactive decision support and data exploration environments. In such settings, users typically pose complex queries, which require considerable amounts of time to produce exact answers, over large parts of the stored data. However, due to exploratory behavior, users can often tolerate small imprecisions in query results, as long as these results are quickly generated and accompanied with accuracy guarantees.

Several studies have demonstrated the applicability of wavelets as a data reduction tool for a variety of database problems. Briefly, the key idea is to first apply the decomposition process over an input data set, thus producing a set of wavelet coefficients. One, then, retains only a subset, composing the *wavelet synopsis*, of the coefficients by performing a thresholding procedure. Clearly, such a lossy compression procedure introduces some error when reconstructing the original data. The bulk of recent work focuses on defining useful metrics that capture this reconstruction error and, further, provide algorithms for constructing optimal synopses given a space constraint.

In a data streaming setting, usually one needs to resort to approximation in order to deal with the high volume and rate of incoming data. Wavelet synopses seem to be an effective summarization technique that can be applied in such a setting as well. Unfortunately, algorithms for constructing wavelet synopses

designed to operate on static disk-resident data cannot be easily extended to process data streams. For example, most of the static algorithms require many passes over the data, whereas, in a streaming context only one-pass algorithms can be applied. In other words, once a data stream item has been processed it cannot be examined again in the future, unless explicitly stored; of course, explicitly storing the entire data stream is not an option.

In this work, we briefly introduce the wavelet decomposition in Section 2. We present our problem formulation for constructing wavelet synopses, discuss the challenges that arise within a data streaming environment and describe our contributions in Section 3. Finally, we conclude our discussion and propose future research directions in Section 4.

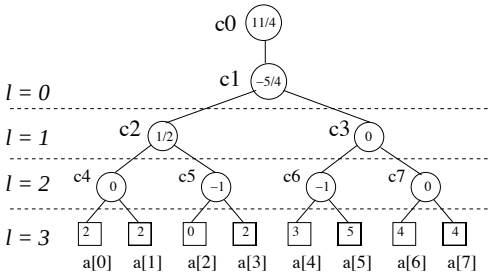
2 Background on Wavelet Decomposition

The wavelet decomposition is a mathematical tool for the hierarchical decomposition of functions with a long history of successful applications in signal and image processing [15]. Let us briefly introduce the wavelet decomposition process through a simple example. Consider the *data vector* $a = [2, 2, 0, 2, 3, 5, 4, 4]$, of domain size $N = 8$. The Haar wavelet decomposition, the simplest of all wavelet decompositions, of a is computed as follows. We first average the values together pairwise to get a new "lower-resolution" representation of the data with the pairwise averages $[\frac{2+2}{2}, \frac{0+2}{2}, \frac{3+5}{2}, \frac{4+4}{2}] = [2, 1, 4, 4]$. This averaging loses some of the information in a. To restore the original a values, we need *detail coefficients*, that capture the missing information. In the Haar decomposition, these detail coefficients are the differences of the (second of the) averaged values from the computed pairwise average. Thus, in our simple example, for the first pair of averaged values, the detail coefficient is 0 since $2 - 2 = 0$, for the second it is -1 since $1 - 2 = -1$. No information is lost in this process – one can reconstruct the eight values of the original data array from the lower-resolution array containing the four averages and the four detail coefficients. We recursively apply this pairwise averaging and differencing process on the lower-resolution array of averages until we reach the overall average, to get the full Haar decomposition, depicted in Figure 1(a). The transform of a is given by $w_a = [11/4, -5/4, 1/2, 0, 0, -1, -1, 0]$, that is, the overall average followed by the detail coefficients in order of increasing resolution. Each entry in w_a, be it a detail or average, is called a *wavelet coefficient*.

A B-term *wavelet synopsis* is simply defined as any subset $\Lambda \subset w_a$ of wavelet coefficients, where usually $B = |\Lambda| \ll N$. Implicitly, all non-stored coefficients are set to 0. Thus, a wavelet synopsis is typically stored by B ⟨coeff-index, coeff-value⟩ pairs.

A useful conceptual tool for visualizing and understanding the hierarchical nature of the Haar decomposition process is the *error tree* structure [12] (shown in Fig. 1(b) for the example array a). Each internal tree node c_i corresponds to a wavelet coefficient (with the root node c_0 being the overall average), and leaf nodes $a[i]$ correspond to the original data-array entries. This view allows us to

Resolution	Averages	Detail Coefficients
3	[2, 2, 0, 2, 3, 5, 4, 4]	–––––
2	[2, 1, 4, 4]	[0, –1, –1, 0]
1	[3/2, 4]	[1/2, 0]
0	[11/4]	[–5/4]



(a) **(b)**

Fig. 1. Example error-tree structure for the example array a.

see that the reconstruction of any $a[i]$ depends only on the $\log N + 1$ coefficients in the path between the root and $a[i]$. Without going into detail, observe that $a[5]$ can be reconstructed by adding or subtracting coefficients in the path from the root down to $a[5]$, depending on whether we descend to a left or right child respectively; i.e., $a[5] = c_0 - c_1 + c_3 - c_6 \Leftrightarrow 5 = \frac{11}{4} - \left(-\frac{5}{4}\right) + 0 - (-1)$. Similarly, notice that the value of a wavelet coefficient only depends on a subset of the original values, depending on the height of the tree they belong to; e.g., the value of coefficient c_5 depends only on the values $a[2]$ and $a[3]$.

Intuitively, wavelet coefficients towards the root of the error tree carry a higher weight in the reconstruction of the original data values. To equalize the importance of all coefficients, a common normalization scheme is to scale the coefficient values at level l by a factor of $\sqrt{N/2^l}$. Letting c_i^* denote the normalized coefficient values, this fact has two important consequences: (1) The *energy* (a.k.a., the L_2 norm) of the a vector is preserved in the wavelet domain, that is, $||a||_2^2 = \sum_i a[i]^2 = \sum_i (c_i^*)^2$ (by Parseval's theorem); and, (2) Retaining the B largest coefficients in terms of *absolute normalized value* gives the (provably) optimal B-term wavelet synopsis in terms of Sum-Squared-Error (SSE) in the data reconstruction (for a given budget of coefficients B) [15]. More formally, assuming a synopsis Λ and denoting by $\widetilde{a}$ the vector of reconstructed data values, the SSE is defined as $\sum_i (a[i] - \widetilde{a}[i])^2 = \sum_{\forall c_j \notin \Lambda} (c_j^*)^2$, where the latter equation is due to Parseval's theorem. In other words, SSE is equal to the sum of squared normalized values of the non-stored coefficients, hence the previous observation.

3 Constructing Optimal Wavelet Synopses

In this section we present a generic problem formulation for constructing optimal wavelet synopses. To this end we distinguish among various error metrics and also differentiate on static (disk-resident) and streaming data. Further, we explicitly relate the contributions of our current work with respect to the aforementioned problem formulation.

A *B-term optimal wavelet synopsis* is a wavelet synopsis that minimizes some aggregate reconstruction error metric under a space constraint of B coefficients — therefore, its construction depends on the definition of such an error metric.

Minimizing weighted L_p norm of point errors. Given a wavelet synopsis Λ of some data vector a, let $\mathrm{err}(i)$ denote the *point error*, that is, the reconstruction error for the i-th data value $a[i]$. In Section 2 we considered the point to be the absolute error $\mathrm{err}_{abs}(i) = |a[i] - \widetilde{a}[i]|$ and further applied the L_2 norm to aggregate across all data values, leading to the SSE error metric. Finding the optimal wavelet synopsis for SSE is quite trivial, as discussed. However, the extension to other point errors, such as, for example, the relative error (with sanity bound s) $\mathrm{err}_{rel}(i) = \frac{|a[i]-\widetilde{a}[i]|}{\max\{s,a[i]\}}$ and using other norms, such as the maximum L_∞ norm, to aggregate individual data reconstruction errors, is not as straightforward.

Let $\mathrm{err}(i)$ denote the i-th point error and w_i denote a weight (or, importance) assigned to this error. Using a weighted L_p norm for aggregation we obtain the following generic error metric: $\sum_i w_i \cdot (\mathrm{err}(i))^p$. Unfortunately, since Parseval's theorem can only be applied in the unweighted L_2 norm of absolute errors, no easy to process rewriting of arbitrary aggregate error metrics exists.

There are two approaches to constructing an optimal wavelet synopsis for general weighted aggregated point errors. The first approach, used in [16] for the weighted L_2 error, tries to incorporate the error metric in the decomposition process. The decomposition step for obtaining the average coefficient changes to a weighted average, that is, for two values a, b we obtain $\frac{w_a a + w_b b}{2}$, where weights w_a, w_b can be constructed from the given reconstruction error weights. This approach leads to a different, Haar-like, decomposition in which the SSE metric is exactly the weighted L_2 error metric measured in the conventional Haar decomposition. Therefore, the construction of the optimal, under weighted L_2 norm, synopsis problem translates to the conventional SSE minimization problem.

The second approach, such as the one taken in [3, 4, 14, 6], is the design of algorithms that incorporate the error metric in their operation by exploiting the error tree structure. In short, due to the distributive nature of error metrics, the algorithms solve a dynamic programming recurrence, where the optimal error incurred at a node i in the error tree (for a specified space budget and for a specified set of ancestor nodes retained in the synopsis) depends on the optimal errors incurred at the two children nodes $2i, 2i+1$. The choice to be made involves distributing available space to children nodes and deciding whether to include node i in the set of retained nodes, or not.

Recently [7], it has been observed that restricting the retained synopsis values to the actual decomposition values is suboptimal for other than SSE error metrics. Indeed, consider the case where just one coefficient, the average, is to be maintained in the synopsis. In the case of an SSE-optimal synopsis the optimal value would be the value in the original decomposition, that is, the average. However, in the case of the maximum absolute error metric the optimal value would rather be $(\min + \max)/2$, where min and max are the minimum and maximum values, respectively, in the original data. In light of this observation, one

has to construct a synopsis by searching not only for the best coefficients to choose, but also for their optimal values. The term used for this more generic and computationally harder optimization problem is the construction of optimal *unrestricted wavelet synopses*.

Extending results to multi-dimensional data sets is not straightforward, as it usually requires the design of external memory algorithms. Our work in [9] presents I/O efficient algorithms for constructing SSE optimal wavelet synopses for massive multi-dimensional data sets. In brief, the main idea is to put into memory a part of the data set such that when the wavelet decomposition is performed on this data, we obtain an as large as possible set of finalized coefficient values. The wavelet decomposition of the in-memory data values is performed efficiently by the SHIFT and SPLIT operations, that intuitively: (i) shift the indices of the detail coefficients to their corresponding indices in the final decomposed data set; and (ii) split the energy of the average coefficients to properly update some already calculated coefficients.

Minimizing weighted L_p norm of range-sum errors. For this case we define the *range-sum error*, denoted by $\mathrm{err}(i:j)$, as the summation of reconstruction errors for data values $a[i]$ through $a[j]$: $\mathrm{err}(i:j) = \sum_{k=i}^{j} \mathrm{err}(k)$. Similar to the case of point errors, one can use weighted L_p norms to aggregate across all $N(N+1)/2$ range-sum errors. Further, the first approach for finding a point error optimal synopsis, described previously, apply to the case of range-sum error optimal synopses as well. The work in [11] operates on the prefix-sum array of a and show that one has to follow a similar to the SSE-optimal synopses thresholding procedure for the case of unweighted L_2 aggregation of range-sum absolute errors. Unfortunately, the second approach of incorporating the error metric in the synopsis construction algorithm cannot be directly applied, since no nice distributive property for aggregating range-sum errors can be exploited.

3.1 Streaming Wavelet Synopses

A data streaming environment introduces resource restrictions to conventional static data processing algorithms, due to the high volumes and rates associated with incoming data. Namely: (i) there is not enough space to store the entire stream, as it can be of potentially unbounded size, and thus, data stream items can only be seen once; (ii) data stream items must be processed quickly in real time; and (iii) queries over data streams are of persistent nature and must be continuously and, most importantly, quickly evaluated. Under these restrictions, data stream processing algorithms must have small space requirements and exhibit fast per-item processing and querying time — here, small and quickly should be read as poly-logarithmic to data stream size.

In our context, we are to construct and maintain the optimal wavelet synopsis of a data vector a whose values are continuously updated by the data stream. There are two conceptually different ways to model [5, 13] how the data stream updates the values of a: (i) the *time series model*, where data stream items are appended to the data vector a, that is, the i-th data stream item is the value

$a[i]$; and (ii) the *turnstile model*, where data stream items update the data vector a, that is, each data stream item (i, u) is an update for one of the data values, implying that $a^{new}[i] \leftarrow a^{old}[i] + u$.

Time Series Model. In this data stream model, since the data stream items are appended at the end of the data vector a, only those coefficients, termed the wavelet *fringe*, in the path from the root down to the most recently appended data value change. This means that the bulk of wavelet coefficients (except for the logarithmically small subset that lies in the fringe) have a data value that it is not going to be affected by subsequent data stream items. For a B-term SSE optimal synopsis, this observation leads to a very simple algorithm [5]: maintain the B highest in absolute normalized value coefficients among those whose value is finalized and additionally keep all the coefficients in the fringe. Once a fringe coefficient is finalized the algorithm simply needs to compare its value with the B stored values and construct the new set of stored values by either dropping the coefficient at hand or the smallest one in the stored set. However, in the case of arbitrary error metrics no algorithm that produces an optimal synopsis exists, to the best of our knowledge. The work in [10] provides with a heuristic as to which coefficients to maintain for a maximum (relative or absolute) error optimal synopsis in the time series model: each time a coefficient needs to be dropped, the one which leads to the smallest increase in error is greedily picked.

For the problem of maintaining SSE optimal wavelet synopses, our work in [9] introduces some interesting results. In data streaming applications, as also argued in [2], it is often more appropriate to keep update times small to accommodate for multiple bursty streams, rather that try to save on memory footprint. To this end, the SHIFT/SPLIT operations defined in [9] allow for a trade-off between per-item processing time and available space for maintaining streaming wavelet synopses. Further, in [9] we present the first time and space requirements results for maintaining wavelet synopses over a multi-dimensional time series data stream. All results are provided for both forms of multi-dimensional wavelet decomposition, standard and non-standard [15].

Turnstile Model. The turnstile model is more general in that it allows arbitrary updates to the data vector, and thus, potentially any wavelet coefficient can be affected by a data stream item. This makes keeping track of wavelet coefficients a very hard task, let alone constructing an optimal synopsis. The work in [5] uses a sketch [1] as a means of (probabilistically) maintaining the energy/magnitude of the data vector a. Then, one can estimate any wavelet coefficient by multiplying the energy of the data vector with that of the corresponding wavelet basis vector, as long as the angle among the two vectors is sufficiently large. Constructing an optimal in terms of SSE synopsis, however, requires super-linear in N time. To make matters worse, no results exist for other error metrics.

Our work in [2] deals with maintaining SSE optimal synopses under this more general model (where sketching techniques are the only option) and offers significant time improvements over previous approaches. The crux of our work lies in two novel technical ideas. First, our algorithms work entirely in the wavelet domain: instead of maintaining a sketch over a data vector we choose

to sketch its wavelet decomposition. This is possible as a single data stream update item can be translated to only poly-logarithmically more update items in the wavelet domain. Second and most importantly, our algorithms employ a novel hierarchical group organization of wavelet coefficients to accommodate for efficient binary-search-like identification of high in absolute normalized value coefficients. In addition, a trade-off between query time and update time is established, by varying the hierarchical structure of groups, allowing the right balance to be found for specific data stream scenarios. The algorithms presented in [2] can easily scale to large domain sizes and, further, can be applied to multi-dimensional data streams for both decomposition forms.

4 Conclusions and Future Work Directions

In this work we have presented a problem formulation for constructing wavelet synopses, general enough to embody the majority of existing work in this area. Further, we have explicitly illustrated the contributions of our current work and described, in context, how it relates to the general formulation. From our discussion one can easily deduce that many interesting and challenging issues remain open for constructing optimal wavelet synopses, especially in a data streaming environment. Our future work will try to address some of these.

In particular, when aggregating point errors, the approach of incorporating the desired minimization metric into the wavelet decomposition seems to be the most promising one, as choosing the coefficients can be done in a similar to the SSE minimization process. Further, such an approach can then be easily adapted to operate over data streams for both models. However, similar results for other L_p norms, including minimizing for the maximum error (L_∞), do not exist. It would be interesting to see whether modified Haar wavelet bases, or even other wavelet bases, are suitable for this task.

Optimizing for arbitrary workloads, such as those that include range-sum queries, seem more useful than simply optimizing for point query workloads. However, as also discussed in [11], optimizing for arbitrary workloads seems to be a difficult task. Perhaps, optimizing for a simpler case, such as that of a workload containing just dyadic range-sum queries, can provide some nice heuristics for arbitrary workloads.

Finally, another interesting issue to consider would be devising techniques for space-efficient compression of wavelet synopses. As recently suggested in [8], adaptive quantization can be applied to the coefficient values, and even some clever indexing can be employed to reduce the overhead of identifying retained wavelet coefficients.

References

1. N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings ACM symposium on Theory of computing (STOC)*, 1996.

2. G. Cormode, M. Garofalakis, and D. Sacharidis. Fast approximate wavelet tracking on streams. To appear in *International Conference on Extending Database Technology (EDBT)*, 2006.
3. M. Garofalakis and P. B. Gibbons. Wavelet synopses with error guarantees. In *Proceedings ACM International Conference on Management of Data (SIGMOD)*, 2002.
4. M. Garofalakis and A. Kumar. Deterministic wavelet thresholding for maximum-error metrics. In *Proceedings ACM Principles of Database Systems (PODS)*, 2004.
5. A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *Proceedings International Conference on Very Large Data Bases (VLDB)*, 2001.
6. S. Guha. Space efficiency in synopsis construction algorithms. In *Proceedings International Conference on Very Large Data Bases (VLDB)*, 2005.
7. S. Guha and B. Harb. Wavelet synopsis for data streams: minimizing non-euclidean error. In *Proceedings ACM International Conference on Knowledge Discovery in Data Mining (SIGKDD)*, 2005.
8. S. Guha and B. Harb. Approximation Algorithms for Wavelet Transform Coding of Data Streams. In *Proceedings ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.
9. M. Jahangiri, D. Sacharidis, and C. Shahabi. Shift-Split: I/O efficient maintenance of wavelet-transformed multidimensional data. In *Proceedings ACM International Conference on Management of Data (SIGMOD)*, 2005.
10. P. Karras and N. Mamoulis. One-pass wavelet synopses for maximum-error metrics. In *Proceedings International Conference on Very Large Data Bases (VLDB)*, 2005.
11. Y. Matias and D. Urieli. On the optimality of the greedy heuristic in wavelet synopses for range queries. Technical Report TR-TAU, 2005.
12. Y. Matias, J. S. Vitter, and M. Wang. Wavelet-based histograms for selectivity estimation. In *Proceedings ACM International Conference on Management of Data (SIGMOD)*, 1998.
13. S. Muthukrishnan. Data streams: algorithms and applications. In *Proceedings ACM Symposium of Discrete Algorithms (SODA)*, 2003.
14. S. Muthukrishnan. Subquadratic algorithms for workload-aware haar wavelet synopses. In *Proceedings Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, 2005.
15. E. J. Stollnitz, T. D. Derose, and D. H. Salesin. *Wavelets for computer graphics: theory and applications*. Morgan Kaufmann Publishers Inc., 1996.
16. D. Urieli and Y. Matias. Optimal workload-based weighted wavelet synopses. In *Proceedings International Conference on Database Theory (ICDT)*, 2005.