XWAVE: Optimal and Approximate Extended Wavelets for Streaming Data

Sudipto Guha
University of Pennsylvania
sudipto@cis.upenn.edu

Chulyun Kim
Seoul National University
cykim@kdd.snu.ac.kr

Kyuseok Shim
Seoul National University
shim@ee.snu.ac.kr

Abstract

Wavelet synopses have been found to be of interest in query optimization and approximate query answering. Recently, extended wavelets were proposed by Deligiannakis and Roussopoulos for data sets containing multiple measures. Extended wavelets optimize the storage utilization by attempting to store the same wavelet coefficient across different measures. This reduces the bookkeeping overhead and more coefficients can be stored. An optimal algorithm for minimizing the error in representation and an approximation algorithm for the complementary problem was provided.

However, both their algorithms take linear space. Synopsis structures are often used in environments where space is at a premium and the data arrives as a continuous stream which is too expensive to store. In this paper, we give algorithms for extended wavelets which are space sensitive, i.e., use space which is dependent on the size of the synopsis (and at most on the logarithm of the total data) and operates in a streaming fashion. We present better optimal algorithms based on dynamic programming and a near optimal approximate greedy algorithm. We also demonstrate the performance benefits of our algorithms compared to previous ones through experiments on real-life and synthetic data sets.

1 Introduction

Approximate query processing has recently emerged as a viable solution for dealing with the huge amounts of data, the high query complexities, and the increasingly stringent response-time requirements that characterize decision support systems (DSS) applications.

Due to the exploratory nature of many DSS applications, in several scenarios such as ad-hoc mining or dealing with remote data [8, 1] approximate answers obtained from synopses suffice. In DSS applications, databases with multiple measures are common. For example, market basket database may include information on the revenue, the quantity of being sold and the profits. One natural and widely used tool for synopses with multiple measures is approximate Wavelet representation.

Traditionally, wavelet approximation methods in this multi-measure scenario used either decomposition on individual dimensions, or treated the data as a vector and applied a multidimensional decomposition. As pointed out by Deligiannakis and Roussopoulos in [3], these methods may result in suboptimal solutions. It is not hard to see that the former may store the same coordinate for more than one measure – which stores the coordinate of the coefficient multiple times and wastes space. The latter on the other hand, may be forced to store very small number of coordinate values of which only a few coefficients might help reduce significantly the error and not be effective as well. To remedy this, extended wavelets were proposed in [3]. This problem seeks to optimize the storage utilization by attempting to store the same wavelet coefficient across different measures, thereby eliminating the bookkeeping overhead for one (or possibly, more) of them. They gave an optimal algorithm for the sum of squared error between the representation of the data achieved by the synopsis and the input. They also gave a faster 2-approximation algorithm for the problem of maximizing the sum of weighted squares of the representation, termed as "benefit". The benefit and error add up to weighted sum of squares of the input coefficients and is therefore fixed, thus the problems can be thought of as"complimentary". They also demonstrated that extended wavelets achieve better estimation quality compared to multidimensional wavelets in several cases.

Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004

However, there are two problems with the proposed solutions. (i) Both their algorithms require linear space. Synopsis structures are frequently used in environments where space is at a premium and the data arrives as a continuous stream which is too expensive to store. Thus, linear space algorithms are not desirable in such scenarios. (ii) An approximation algorithm for maximizing the benefit does not give a good approximation algorithm for minimizing error, e.g., suppose the optimum solution had benefit 99 and error 1. Suppose a 2-approximation of the benefit achieved a benefit of 50 (which is more than $\frac{99}{2}$) — but the error of this solution is 50 as well, 50 times the optimum error.

1.1 Our contributions

We make the following contributions:

- To address the problem of linear space, we present optimal algorithms for extended wavelets which are space sensitive, i.e., use space which is dependent on the size of the synopsis (and at most on the logarithm of the total data size) and operates in a streaming fashion.
- For the problem of no guarantee on error by the previous approximation algorithm, we give an algorithm that has error *less than or equal* to the error of the optimum solution (therefore at least as much benefit), but relaxes the space bound to store a few extra coefficients (at most as many as the number of different measures).
- We also demonstrate how to adapt all the above algorithms to the context of streaming data (which connects to the linear space requirement of previous work), i.e., given multidimensional points as a stream we construct the coefficients on the fly as well as maintain the synopsis. This is particularly of use in modeling time series data.
- Through experiments on real-life and synthetic data sets, we demonstrate that our proposed algorithms have significant performance benefits while requiring much less space.

We would like to mention that if the space bound is indeed *strict*, we can give a different $(1 + \epsilon)$-approximation algorithm for the optimum error preserving the space bound. In this process, we show a *non-trivial connection* between extended wavelets and histograms similar to the V-Optimal objective. The complexity of the algorithm is asymptotically the same as the approximation algorithm presented here. However, due to space limitations, this connection cannot be described in this paper. It can be found in [7]. Furthermore, the algorithm that arises from the connection to histograms is somewhat theoretical and significantly complicated to implement, which we relegate to future work.

1.2 Organization

The paper is organized as follows. In the next section, we present related work. In Section 3, we introduce preliminary definitions and the problem of constructing extended wavelet on databases with multiple measures. In Section 4, we introduce improved optimal algorithms. We then present the approximation algorithm in Section 5. Section 6 discusses how to adapt the extended wavelet to stream. In Section 7, we present experimental results. Finally, we make concluding remarks in Section 8. Due to the lack of space, we are unable to present any proofs of the lemmas and theorems. They can be found in [7].

2 Related Work

Several approximation techniques using small summary have been developed for selectivity estimation and approximate query answering. These techniques include histograms [14, 15, 9, 13], wavelets [10, 2] and sampling [4, 20].

Wavelet-based approaches provide a mathematical tool for the hierarchical decomposition of functions, with a long history of successful applications in image processing [12, 16]. In [2, 10, 16], they demonstrated that wavelets can be accurate even in high-dimensional datasets. Recent studies have also demonstrated the applicability of wavelets in selectivity estimation [10], answering range-sum aggregates queries over data cubes [19, 18], approximate query processing [2] and data streams [11, 5].

3 Preliminaries

Wavelets, particularly Haar wavelets, provide useful tools for multi-resolution summarization. In context of databases they have been found to be of interest in query optimization, approximate query answering, and similarity estimation. We review the definition of wavelets before discussing our problem.

3.1 Wavelets

We consider signals indexed on $\{1, \ldots, N\}$, where N is a power of 2. Given a sequence of N numbers $\mathbf{X} = x_1, \ldots, x_N$, which can thought of belonging to the Euclidean space $\Re^N$, we can represent the sequence as a linear combination $\sum_{i=1}^{N} x_i \mathbf{u}_i$ where $\mathbf{u}_i$ is the N-dimensional vector where the i-th coordinate is set to 1 and all other coordinates are 0.

Definition 3.1 The function that equals 1 on set S and zero elsewhere is denoted by $\mathbf{\Gamma}(S)$. A (Haar) wavelet is a function $\mathbf{\Psi}$ on $[1, N]$ of one of the following forms:

- $\frac{1}{\sqrt{N}}\mathbf{\Gamma}([1, N])$