

# 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

---

*Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment.*

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

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.





















