Wavelet-based histograms for selectivity estimation: a systematic study

November 2, 2001

Cristelle Barillon, Yossi Matias, Min Wang

1 Introduction

Query optimization is an integral part of database systems. Matias, Vitter and Wang in [1], described a technique based on wavelet transform for building efficient histograms of the underlying data distribution. It proved to offer substantial improvement in accuracy over previous approaches.

On the other hand, wavelets are a mathematical tool used in miscellaneous applications, and most successfully in signal processing, for compression, denoizing etc... of a signal (see [2] for details). They provide a hierarchical representation of fonctions according to their frequencies and localization in time or space.

In [1], the authors focused their attention to finalizing this new technique, and thus used the simpliest wavelet: the Haar wavelet. Our work here consisted in a systematic study of more sophisticated wavelets in order to understand if more significant improvements could be obtained. The different of wavelets that were tested are:

- 8 Orthogonal (Daubechies) Wavelets: the simpliest of the compactly-supported orthogonal wavelets is the Haar transform, each of the Daubechies Wavelet is a refienement, providing smoother approximation, avoiding the "blocking effect" of the haar approximation.
- 4 Coifflet Wavelets : a different kind of orthogonal wavelets
- 10 Biorthogonal Spline Wavelets: those are very elaborated and present many interesting mathematical properties. They are supposed to provide good synthesis (interpolation after thresholding) .
- 3 Interpolating (Deslauriers-Dubuc) wavelets: they are very simple wavelets, also good for interpolation

The tests were performed on synthetic data sets which are indicative of various real life data. Orthogonal Daubechies wavelets give a good improvement for several kinds of queries and all error measures. Biorthogonal wavelets are also interesting although the improvement they provide is not significant for all error measures. Coifflet and Deslaurier-Dubuc wavelets do not, most of the time, give significant improvement.

2 Datas, query sets

As in [1], " the spread of the value set follows the *cusp-max* distribution with Zipf parameter $z = 1$, the frequency set follows a Zipf distribution with parameter z" varying for 0.1 to 1.5, "and frequencies are randomly assigned to the elements of the value set. The value set size is $n = 500$, the domain size is $N = 4096$, and the relation size is $T = 10^5$".

Query set 1 is the set of one-sided range queries (it corresponds to query A according to the notations of the article). **Query set 2** is the set of equal range queries (it corresponds to query G of the article). **Query set 3** is the set of all possible two-sided range queries (it corresponds to query 5 in the code and query C of the article). **Query set 4** is the set of all possible two-sided range queries with range size Δ (it corresponds to query 6 in the code and query E of the article). **Query set 5** is the set of all possible two-side range queries with range size equal to half of the dimension size (it corresponds to query 7 in the code, and query set E of the article).

We used the same error measures as in [1].

3 Results

Most of the time, the orthogonal Daubechies wavelet with 4 coefficients performs very satisfyingly. Here is on each query type, the error measures compared to the haar wavelet for synthetic datas with Zipf distribution 0.5:

. Query set 1

error norm	haar	Daubechies 4
$L1_{abs}$	1251.091333	227.578981
$L2_{abs}$	1496.834787	289.626637
$L1_{rel}$	0.140706	0.012008
$mod.L1_{rel}$	0.146397	0.037381
$L1_{com}$ $\beta = 100$	11.937757	1.169560
$L2_{com}$ $\beta = 100$	104.345905	4.206917
$L1_{com}$ $\beta = 10$	1.407057	0.120084
$L2_{com}$ $\beta = 10$	17.557885	0.470052

. Query set 2

error norm	haar	Daubechies 4
$L1_{abs}$	48.127489	36.534931
$L2_{abs}$	348.698800	64.385951
$L1_{rel}$	14.763380	18.571439
$mod.L1_{rel}$	38.770321	19.729223
$L1_{com}$ $\beta = 100$	41.673008	31.432501
$L2_{com}$ $\beta = 100$	329.606536	37.209406
$L1_{com}$ $\beta = 10$	18.119517	20.076329
$L2_{com}$ $\beta = 10$	261.367219	24.075499

. Query set 3

error norm	haar	Daubechies 4
$L1_{abs}$	1716.837238	322.674478
$L2_{abs}$	2117.102507	409.491873
$L1_{rel}$	0.244150	0.174231
$mod.L1_{rel}$	91.795916	0.268182
$L1_{com}$ $\beta = 100$	17.210921	3.599531
$L2_{com}$ $\beta = 100$	67.261953	14.108730
$L1_{com}$ $\beta = 10$	1.801908	0.494316
$L2_{com}$ $\beta = 10$	19.270774	5.220866

. Query set 4 with $\Delta = 10$

error norm	haar	Daubechies 4
$L1_{abs}$	446.758585	145.681979
$L2_{abs}$	1055.565653	202.499969
$L1_{rel}$	7.027228	14.349136
$mod.L1_{rel}$	234.950665	17.237898
$L1_{com}$ $\beta = 100$	187.584968	76.539022
$L2_{com}$ $\beta = 100$	557.361031	108.372850
$L1_{com}$ $\beta = 10$	25.226143	21.212737
$L2_{com}$ $\beta = 10$	159.744390	60.624457

. Query set 5

error norm	haar	Daubechies 4
$L1_{abs}$	1009.806875	355.298044
$L2_{abs}$	1198.849908	447.987513
$L1_{rel}$	0.020342	0.007189
$mod.L1_{rel}$	0.020668	0.007225
$L1_{com}$ $\beta = 100$	2.034239	0.718866
$L2_{com}$ $\beta = 100$	2.413644	0.910392
$L1_{com}$ $\beta = 10$	0.203424	0.071887
$L2_{com}$ $\beta = 10$	0.241364	0.091039

In figure 1 to 5 we show one of the error measures for differents wavelets as the Zipf distribution changes for $z = 0.1$ to $z = 1.5$ on the 5 query sets. The

wavelets in comparison here are the orthogonal 4 coefficients and 18 coefficients, the 3rd interpolating wavelet and the biorthogonal 4 and 18 coefficients. All of them perform better for low values of z, and only the orthogonal wavelets seem to maintain this performance for high values of z.

In figure 1, the L1 relative error is shown as a function of z for query set 1.

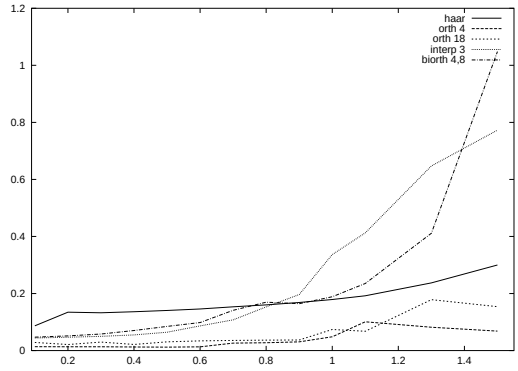

Figure 1: Query set 1

In figure 2, the L1 absolute error is shown as a function of z for query set 2.

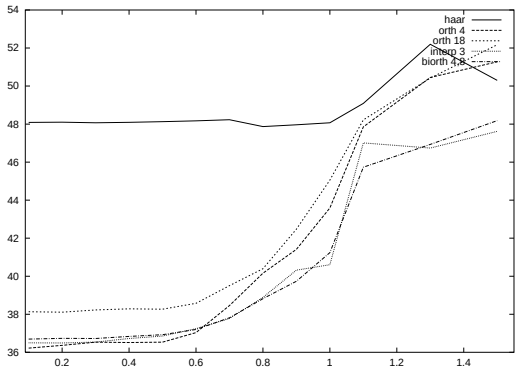

Figure 2: Query set 2

In figure 3, the L1 relative error is shown as a function of z for query set 3.

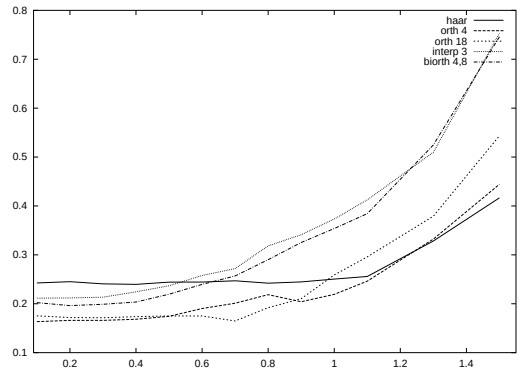

Figure 3: Query set 3

In figure 4, the L1 absolute error is shown as a function of z for query set 4.

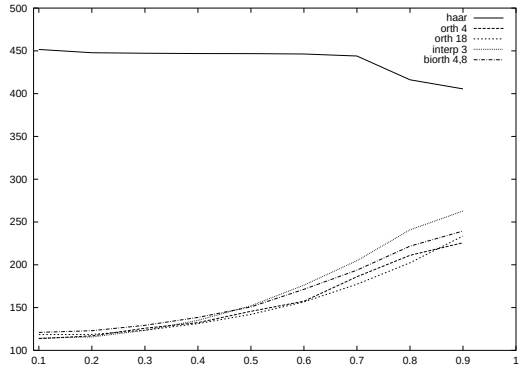

Figure 4: Query set 4 with $\Delta = 10$

In figure 5, the modified L1 error is shown as a function of z for query set 5

.

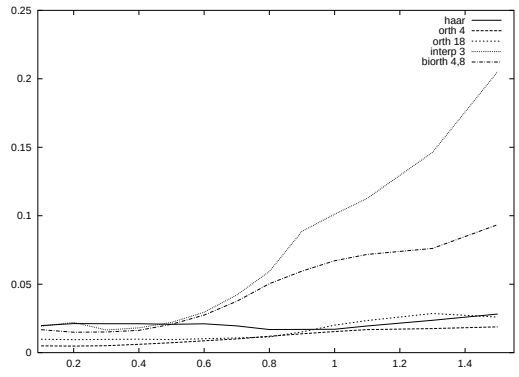

Figure 5: Query set 5

In figure 6, we show the L1 relative error measure for query set 1 for haar, orthogonal 4 and 18 coeeficients, 3rd interpolating and biorthogonal 4,8 wavelets with storage between 21 and 41 coefficient out of 1024 non void values. Hence the storage space is a less essential feature for non haar wavelets, the error measure for those being quite stable contrary to the error corresponding to the haar approximation which decreases frankly as the storage space increases. We avoid the so-called blocking effect.

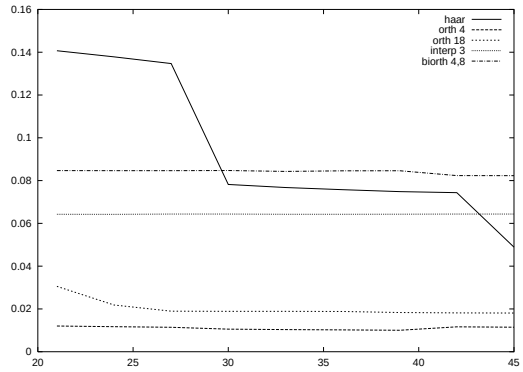

Figure 6: Influence of the storage space on the approximation

As a conclusion, we can say that some improvement is provided by more sophisticated wavelets than the Haar transform. Although this improvement is not as obvious and systematic as desired, a good compromise can be the simple 4 coefficients orthogonal wavelet that turns out to be very effective. Further improvement should be obtained by a better treatment of the edges of the data and the use of wavelet packs.

References

[1] Y. Matias, J.S. Vitter and M. Wang. Wavelet-based histograms for selectivity estimation. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of Data*, pp. 448-459, Seattle, WA, June 1998.

[2] S.Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1998