상대 오차의 최소화를 위한 조화 웨이블릿 기법

함성호° 강성구 이석호
서울대학교 전기컴퓨터공학부
{sungho°, exodus}@db.snu.ac.kr, shlee@cse.snu.ac.kr

Harmonic Wavelet Method for Minimizing Relative Error

Sungho Ham° Seonggoo Kang Sukho Lee
School of Electrical Engineering and Computer Science, Seoul National University

요 약

대용량 데이타에 대한 복잡한 질의 처리가 요구되는 응용에서 빠른 시간 안에 응답을 돌려주기 위해, 데이타를 작은 크기로 근사하여 질의를 처리하는 방법이 연구되고 있다. 빠른 응답을 위해 주어진 저장 공간의 제약 하에서 얼마나 원본 데이타와 유사하게 근사할 수 있는지가 데이타 근사의 성능을 결정한다. 본 논문에서는 데이타 근사에서 유사도의 척도 중 하나인 최대 상대 오차를 줄이기 위하여 Haar 웨이블릿을 변형한 조화 웨이블릿 기법을 제안한다. 조화 웨이블릿은 데이타 변환 과정 중 조화평균을 이용하여 상대 오차 정보를 손쉽게 얻어낼 수 있어 근사 데이타의 상대 오차를 최소화하는 데 적합한 기법이다.

1. 서 론

다양한 데이타 베이스 응용에서 다루는 데이타의 크기가 커짐에 따라 데이타 근사를 이용한 효율적인 근사 질의 처리 방법이 연구되고 있다. 대용량 데이타를 다루는 대표적인 응용인 의사결정 시스템에서는 복잡한 질의의 처리가 요구되어 근사 질의 처리를 이용해 질의 처리시간을 줄여 왔다. 최근 데이타 베이스 응용이 데이타 스트림과 유비쿼터스 컴퓨팅으로 확장되며 실시간으로 발생하는 데이타의 저장과 처리를 위한 데이타 근사의 필요성이 더욱 커지고 있다.

최근 데이타 근사 연구에서 널리 사용되고 있는 웨이블릿(wavelet)은 데이타의 계층적 분해를 가능하게 해주는 수학적 기법이다. 웨이블릿 기법을 사용해 얻어진 데이타의 분해결과를 저장 공간의 제약에 맞도록 요약하여 원본 데이타 대신 사용하게 된다. 요약된 분해결과를 이용하면 작은 저장 공간만을 할당하여 원본과 유사한 근사값을 얻어낼 수 있으며, 질의 처리 시 탐색공간을 줄여 효율적인 질의 처리를 할 수 있다[1,2]. 이렇게 얻어진 근사 데이타는 원본과의 오차를 가지게 되는데 이것을 최소화 하는 것이 데이타 근사의 목표이다.

통상적인 웨이블릿 기반 데이타 근사에서는 Haar 웨이블릿을 사용해 얻어진 데이타 분해 결과를 L_2-norm 오차를 최소화 하도록 요약하는 방법을 사용하였다. 그러나 L_2-norm 오차로는 오차가 원본 데이타에 비해 어느 정도 큰 값인지 판별할 수 없으며, 개별 데이타 값 사이에서 발생하는 오차들의 편차 역시 알 수 없다는 문제가 있다. 이 때문에 최근에는 오차와 원본 데이타 사이의 비율 값의 상한인 최대 상대 오차를 최소화 하기위한 연구가 계속되고 있다[3,4].

또한, 최대 상대 오차의 최소화에서 Haar 웨이블릿의 분해 방법에서 비롯되는 한계를 극복하기 위한 새로운 웨이블릿 기법의 필요성이 제기되고 있다. Harr 웨이블릿은 산술평균을 이용하여 데이타를 분해하는데 이렇게 분해된 결과는 절대적인 값의 차이만을 반영하고 있다. 따라서 절대적인 차이를 나타내는 L_2-norm 오차는 쉽게 최소화[5]할 수 있으나 상대 오차를 최소화하는 데는 적합하지 않은 것이다.

본 논문에서는 데이타 근사결과의 상대 오차를 줄이기 위하여 새로운 웨이블릿 기법인 조화 웨이블릿을 제안한다. 이 기법은 Haar 웨이블릿에서의 산술평균 대신 조화평균을 이용하여 데이타를 분해한다. 데이타 분해시 조화 평균을 이용하면 상대 오차 정보가 얻어져 최대 상대 오차의 최소화를 쉽게 줄일 수 있다. 또한, 조화 웨이블릿 기법은 기존의 Haar 웨이블릿의 질의 처리방법과 오차 최소화 방법에 손쉽게 적용할 수 있다는 장점을 가진다.

본 논문의 구성은 다음과 같다. 2절에서는 질의 처리와 오차의 최소화에 관한 연구들을 설명한다. 3절은 Haar 웨이블릿을 변형한 조화 웨이블릿을 제안한다. 4절에서는 Haar 웨이블릿과 조화 웨이블릿의 성능을 비교하며, 5절에서 결론과 향후 연구 방향을 제시한다.

2. 관련연구

근사 질의 처리의 방법으로는 웨이블릿 이외에 샘플링(sampling), 히스토그램(histogram)을 이용한 방법들이 있다. 그러나 웨이블릿 기법은 이들과 달리 데이타 근사를 위한 수학적인 모델을 제공해 주며 다양한 질의 처리에 효율적인 응용이 가능하다는 장점이 있다[2]. Haar 웨이블릿을 이용해 분해된 데이타를 요약하는 방법에는 L_2-norm을 최소화 하기위한 방법[5]과 최대 상대 오차를 최소화 하기위한 방법[3,4]이 있다. 최대 상대 오차는 정규화된 표준 오차를 통해 확률적(probabilistic)으로 최소화[3]되거나 확정적(deterministic)인 다이나믹 프로그래밍(DP) 알고리즘[4]을 이용하여 최소화될 수 있다.

1) 본 연구는 2005년도 두뇌한국21사업과 정보통신부 대학 IT연구센터(ITRC) 지원을 받아 수행되었습니다.

3. 조화 웨이블릿 기법

본 논문에서 사용할 기호에 대한 의미를 표 1에 설명하였다. 이 후의 내용에서 필요에 따라 기호들을 다시 설명 하도록 한다.

표 1. 사용된 기호들의 의미

기호	의미
N	데이타 배열의 길이
B	가용한 웨이블릿 계수 저장 공간
A	입력 데이타 배열
W_A	A에 대한 웨이블릿 분해 결과
d_i	i번째 데이타
$\hat{d}_i$	요약된 계수로부터 재구성된 i번째 데이타
c_i	i번째 웨이블릿 계수
T_i	c_i를 루트 노드로 하는 하위 오차 나무

3.1 산술평균과 조화평균의 비교

데이타 x, y로 이루어진 데이타 배열 $[x\ y]\ (0 < x \le y)$의 산술평균 $M_A = \frac{x+y}{2}$ 이다. 산술평균은 x,y 각각의 값과 동일한 절대오차 $M_A - x = y - M_A = \frac{y-x}{2}$를 가진다. 임의의 실수 k값 중 M_A보다 작은 것은 y쪽에서 더 큰 오차를 발생시키고, M_A보다 큰 것은 x쪽에서 더 큰 오차를 발생시킨다. 따라서 Haar 웨이블릿의 분해과정에서 사용되는 산술평균은 절대오차 최소화의 관점에서 최적의 선택이 될 수 있다. 그러나, 상대 오차의 경우 M_A와 x사이의 오차 보다 M_A와 y사이의 오차가 더 커지므로 상대 오차를 최소화하기 위해서는 산술평균 대신 x와 y에서 같은 상대 오차를 발생시키는 평균값을 사용해야 한다.

임의의 k $(x \le k \le y)$와 x사이에서는 $\frac{k-x}{x}$, y사이에서는 $\frac{k-y}{y}$의 상대 오차가 발생한다. $\frac{k-x}{x} = \frac{k-y}{y}$를 k에 대하여 정리해 보면 $k = \frac{2xy}{x+y} = \left(\frac{1/x + 1/y}{2}\right)^{-1}$이 얻어지며, 이 것은 x, y의 조화평균에 해당하는 값이다. 따라서 Haar 웨이블릿의 분해과정에서의 산술평균을 조화평균으로 대신하는 새로운 분해방법으로 상대 오차를 최소화 할 수 있다.

3.2 조화 웨이블릿

조화 웨이블릿은 근사된 데이타와 원본 데이타 간의 상대 오차를 줄이기 위하여 Harr 웨이블릿을 변형한 기법이다. 조화 웨이블릿은 분해과정에서 Harr 웨이블릿에서의 산술평균과 차 대신 조화평균과 상대 오차를 이용한다.

데이타 x, y로 이루어진 길이 2^1인 데이타 배열 $A = [x\ y]$을 분해하려면 두 데이타 값을 짝을 지어 조화평균을 계산하고 조화평균과 데이타값 사이의 상대 오차를 얻어내야 한다. 이때, 조화평균과 데이타 값 사이의 상대 오차는 x, y에서 동일하게 발생한다. A의 분해의 결과는 계산된 조화평균과 상대오차로 이루어진 $W_A = [\frac{2xy}{x+y}\ \ 1 - \frac{2y}{x+y}]$가 된다. 길이가 2^n인 데이타 배열의 경우 전체 데이타 배열에 대한 평균이 구해질 때까지 앞서 설명한 과정을 재귀적으로 n번 반복한다. 분해과정을 모두 끝마친 후 얻어진 값을 웨이블릿 계수(wavelet coefficient)라 한다. 표 2는 길이가 2^2인 데이타 [12 8 6 4]가 웨이블릿 계수 [6.4 0.33 0.2 0.2]로 분해되는 과정이다. 가장 먼저 12와 8이 짝지어져 조화평균 9.6과 상대오차 0.2가 계산되고, 다음으로 6과 4가 짝지어져 4.8과 0.2가 얻어진다. 이 과정에서 얻어진 평균값 9.6과 4.8이 다시 짝지어져 조화평균과 상대 오차가 계산되어 분해과정이 끝나게 된다.

표 2. 조화 웨이블릿의 분해과정

해상도	조화평균	상세계수
2	[12 8 6 4]	-
1	[9.6 4.8]	[0.2 0.2]
0	[6.4]	[0.33]

분해 과정에서 값을 둘 씩 짝을 지어 조화평균을 계산하면 한 단계 낮은 해상도(resolution)의 값을 얻게 된다. 조화평균과 짝지었던 값과의 상대 오차를 상세계수(detailed coefficient)라 한다. 해상도 n의 조화평균 값과 상세계수를 사용하면 해상도 n+1의 조화평균 값을 복원해 낼 수 있다.

오차나무(error tree)를 이용하면 분해결과로부터 데이타를 재구성 할 수 있다. 나무의 내부 노드는 분해의 결과로 얻어진 웨이블릿 계수 c_j들이고 단말 노드는 재구성된 데이타 값 d_i들이다. 그림 1(a)는 [12 8 6 4]의 분해 과정으로 얻어진 오차 나무이다.

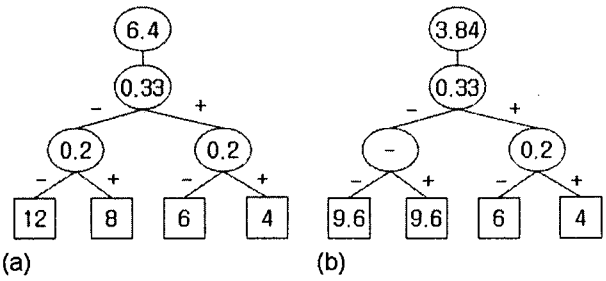

그림 2. 오차나무

오차나무 상의 웨이블릿 계수 c_j들은 루트부터 노드가 위치하는 너비우선탐색 순으로 0 ~ $N-1$의 인덱스를 가지게 된다. 데이타 값 d_i역시 순서대로 0 ~ $N-1$의 인덱스를 가지게 된다. d_i는 c_j의 값으로 부터

$$d_i = c_0 \left\{ \prod_{c_k \in path(d_i) - \{c_0\}} 1/(1 + sign_{ij} c_j) \right\} \quad (1)$$

와 같이 얻어진다. $path(d_i)$는 루트부터 d_i까지의 경로

에 있는 계수들의 집합이고, $sign_{ij}$는 d_i가 c_j의 왼쪽에 위치하면 -1, 오른쪽에 위치하면 $+1$의 값을 갖는다. 예를 들어, $d_3 = c_0/\{(1+c_1)(1-c_3)\} = 6.4/(1.33)(0.8) = 6$ 으로 재구성된다.

3.3 웨이블릿 계수의 요약

분해 과정을 거친 후 얻게 되는 웨이블릿 계수는 원본 데이타와 크기가 동일하며, 정보의 손실이 없어 재구성의 결과로 원본 데이타와 동일한 값을 얻을 수 있다. 저장 공간의 제약을 만족시키기 위해서는 웨이블릿 계수 중 중요한 몇 개만을 선택하고 나머지는 버리는 요약 과정이 필요하다. 버려진 웨이블릿 계수는 0의 값을 가지는 것으로 처리 하여 (1)과 동일한 방법으로 $\hat{d}_i$를 얻어낼 수 있다. 그림 1(b)는 계수 c_1을 버렸을 때의 요약 결과이다. 0.2의 값을 가졌던 계수가 0으로 처리되어 재구성된 데이타 값에서 오차가 발생하였다. 이때 발생하는 최대 상대 오차는 $|(d_0-\hat{d}_0)/d_0| = |(12-9.6)/12| = 0.2$이다.

어떤 계수를 남겨 놓는가에 따라 근사된 데이타의 오차가 결정되므로, 정해진 오차의 척도를 최소화 할 수 있는 계수를 선택하는 것이 중요하다. 확정적인 DP 알고리즘[4]에서 데이타를 재구성하는 부분을 (1)로 수정하면 조화 웨이블릿에의 적용이 가능해진다. 수정된 알고리즘을 이용하면 조화 웨이블릿에서 최대 상대 오차에 대한 최적의 계수들을 선택할 수 있다.

DP에서 채워나가는 테이블 요소는 $M[j,b,S]$로 표현된다. $M[j,b,S]$에는 하위나무 T_j에 대해 공간 b가 할당되어있고, c_j의 조상노드들 중 선택된 웨이블릿 계수의 집합이 S일 때 최대 상대 오차의 최소값이 저장된다. 테이블 요소 $M[0,B,\varnothing]$에 전체 오차나무에 계수저장 공간 B가 주어졌을 때 최대 상대 오차의 최소값이 저장된다.

편의상 $d_{i-N} = c_i$ $(i = N, N+1 \ldots 2N-1)$로 정의하면, 단말노드$(N \le j \le 2N-1)$ 인 경우 테이블 요소는

$$M[j,0,S] = \left| \frac{d_{j-N} - \hat{d}_{j-N}}{d_{j-N}} \right|$$

로 정의된다. 여기서, 재구성된 데이타인 $\hat{d}_i$는 d_i의 부모 중 선택된 계수의 집합을 s_i라 할 때

$$\hat{d}_i = \begin{cases} \prod_{c_k \in s_i} 1/(1+sign_{ij}c_j) & (c_0 \notin s_i) \\ c_0 \left\{ \prod_{c_k \in s_i - \{c_0\}} 1/(1+sign_{ij}c_j) \right\} & (c_0 \in s_i) \end{cases}$$

와 같이 정의된다.

내부 노드$(0 \le j \le N-1)$의 경우 c_j를 버렸을 때

$$\min_{0 \le b' \le b} \max \begin{Bmatrix} M[2j,b',S] \\ M[2j+1,b-b',S] \end{Bmatrix}$$

, c_j를 선택했을 때

$$\min_{0 \le b' \le b} \max \begin{Bmatrix} M[2j,b',S \cup c_j] \\ M[2j+1,b-b'-1,S \cup c_j] \end{Bmatrix}$$

로 정의된다. 알고리즘의 시간 복잡도는 $O(N^2 B log B)$ 공간 복잡도는 $O(N^2 B)$이다.

4. 실험결과

Haar 웨이블릿과 조화 웨이블릿을 사용해 분해한 웨이블릿 계수에서 정해진 수만큼의 계수를 선택했을 때 최대 상대 오차의 최소값을 측정하였다. 실험에 사용한 데이타의 길이는 256이고, 데이타의 값들은 가우스 분포(gaussian distribution)를 따른다.

그림 2에서 요약된 계수들이 원본 데이타의 특성을 어느 정도 반영하게 되는 시점 이후에는 조화 웨이블릿이 약 20%가량 작은 최대 상대 오차를 발생시키는 것을 확인 할 수 있다.

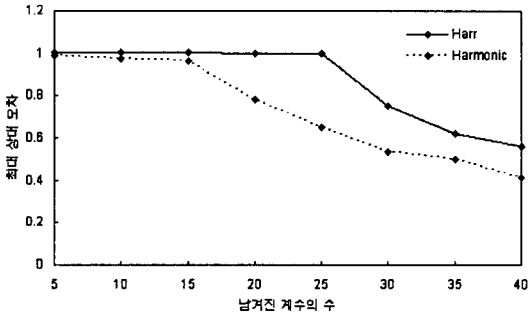

그림 2. 남겨진 계수의 수에 따른 상대 오차

5. 결론

본 논문에서는 근사 질의 처리에서 근사된 데이타의 상대 오차를 줄이기 위한 조화 웨이블릿을 제안하였다. 조화 웨이블릿은 조화 평균을 이용하여 상대 오차를 줄일 수 있으며, 기존의 Haar 웨이블릿을 기반으로 한 연구들에 쉽게 적용이 가능하다.

추후 연구에서는 복수개의 애트리뷰트를 갖는 테이블에 대한 질의처리를 위하여 다차원으로 확장하고, 질의 연산자에 대한 연산 방법을 정의할 필요가 있다.

참고문헌

[1] Jeffery Scott Vitter and Min Wang. "Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets". ACM SIGMOD, Philadelphia, Pennsylvania, May 1999.

[2] Kaushik Chakrabarti, Minos Garofalakis, Rajeev Rastogi, and Kyuseok Shim. "Approximate Query Processing Using Wavelets". VLDB, pp 111-122, September 2000.

[3]Minos Garofalakis, Phillip B. Gibbons, "Wavelet Synopses with Error Guarantees",ACM SIGMOD , pp 476-487, June 2002.

[4]Minos Garofalakis, Amit Kumar, "Deterministic Wavelet Thresholding for Maximum-Error Metrics",ACM SIGMOD, pp166-176, June 2004.

[5] E.J Stollnitz, T.D.DeRose, and D.H. Salesin. "Wavelets for Computer Graphics", Morgan Kaufmann Publishers, Inc., 1996.