시계열 데이타베이스에서의 효율적인 유사 검색을 위한 Polar Wavelet 기법

이범기° 강성구 이상준 이석호
서울대학교 전기컴퓨터공학부
{bklee°, exodus, freude}@db.snu.ac.kr, shlee@cse.snu.ac.kr

Polar Wavelet Method for Efficient Similarity Search in Time Series Databases

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

요 약

유클리드 거리에 기반하여 유사한 시퀀스 검색을 하는 기법들은 각 시퀀스에서 특징을 추출하여 차원을 감소시킨 후, R-tree 같은 다차원 인덱싱 기법을 사용하여 검색을 수행한다. 본 논문에서는 시계열 데이타베이스에서의 유사 검색 성능 향상을 위한 새로운 특징 추출 기법인 Polar Wavelet 기법을 제안한다. 이 기법은 유사 검색시 후보 시퀀스의 개수를 줄임으로써 검색 성능을 향상시킬 수 있고, 특징 추출을 위해 시퀀스의 길이를 2^n으로 만들 필요가 없는 장점을 갖고 있다.

1. 서 론

시계열 데이타베이스란 시간에 따라 변화되는 객체의 일련의 값들로 구성된 데이타 시퀀스들의 집합으로 컴퓨터에 저장되어 있는 데이타의 많은 부분을 차지하고 있다. 이러한 시계열 데이타베이스에서 유사한 시퀀스를 검색하는 것은 데이타 마이닝이나 데이타 웨어하우스 같은 분야에서 매우 중요한 역할을 하고 있다[1].

유사 검색이란 주어진 질의 시퀀스와 변화의 패턴이 유사한 시퀀스들을 시계열 데이타베이스에서 찾아내는 것이다. 길이가 동일한 두개의 시퀀스 $\vec{x} = (x_1, x_2, \ldots, x_n)$, $\vec{y} = (y_1, y_2, \ldots, y_n)$의 유사도를 나타내는 척도로는 다음과 같은 유클리드 거리(Euclidean distance)가 주로 사용되며, 유클리드 거리가 주어진 값 ε보다 작을 경우 두 시퀀스는 유사하다고 말한다.

$$D(\vec{x}, \vec{y}) = \left(\sum_{i=0}^{n-1} |x_i - y_i|^2 \right)^{\frac{1}{2}}$$

본 논문에서는 대상이 되는 후보 시퀀스를 줄임으로써 검색 성능을 향상시킬 수 있는 새로운 특징 추출 기법인 Polar Wavelet을 제안한다. 이 기법은 특징을 추출하기 위해 극좌표(polar coordinates)를 사용하며, 차원을 감소시키면서도 착오 누락(false dismissals)이 없음을 보장한다. 또한, 기존의 Wavelet을 이용한 특징 추출 기법과는 달리 특징을 추출하기 위해 시퀀스의 길이를 2^n으로 만들 필요가 없다. 대상이 되는 시계열 데이타베이스는 각 시퀀스의 모든 데이타 값이 $\pi/4$보다 큰 경우로 한정하며, 시퀀스 길이가 동일한 전체 매칭(whole matching) 유사 검색에 적용할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서 시계열 데이타베이스에서의 유사검색 및 특징 추출 기법과 관련된 연구에 대해 살펴보고, 3절에서 본 논문에서 제안하는 기법에 대해 설명하며, 4절에서 결론을 맺는다.

2. 관련연구

시계열 데이타베이스에서는 효율적인 검색을 위해 R-tree와 같은 다차원 인덱싱 기법 (multi-dimensional indexing methods)을 사용한다. 이때 사용되는 시퀀스가 대개 길이가 긴 고차원이기 때문에 문제 (dimensionality curse)가 발생한다. 즉 인덱스 차원이 높아질수록 순차적 검색보다도 성능이 나빠지는 경우가 발생할 수 있다[1,2,3]. 그러므로 대부분의 유사 검색 기법에서는 차원을 감소시키기 위해 데이타 시퀀스에서 시퀀스 내 데이타 개수보다 적은 k개의 특징을 추출하여 그 특징을 k차원 공간으로 사상하는 특징 추출 기법을 사용한다[1,2,3,4,5].

효율적인 특징 추출 기법으로는 DFT[2,4], DWT[1,6], PAA[5] 등이 제안되어 왔다.

3. Polar Wavelet 기법

3.1 Polar Wavelet

본 논문에서 제안하는 Polar Wavelet 기법은 그림 1과 같은 극좌표 방법을 이용한다. Harr Wavelet은 두 수 사이의 평균과 평균으로부터의 거리를 특징으로 추출하지만, Polar Wavelet은

두 수로부터 만들어지는 극좌표상의 반지름과 cosθ 값을 특징으로 추출한다.

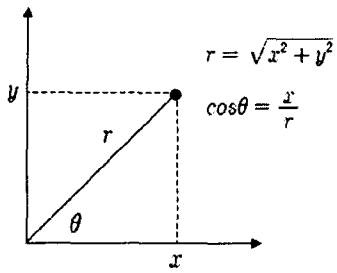

그림 1. 극좌표

분해는 시퀀스로부터 특징을 추출하는 과정으로 시퀀스 [3 4 5 6]의 단계별 분해 과정은 표 1과 같다. 예를 들어 3과 4가 만드는 반지름 값은 $\sqrt{3^2+4^2}$이므로 5가 되며, cosθ 값은 3/5이 된다. 분해 과정 후 추출하는 특징의 개수는 인덱스 구성에 따라 달라지겠지만 4개를 추출할 경우 [$5\sqrt{5}$ $1/\sqrt{5}$ 3/5 3/5]이 된다.

표 1. Polar Wavelet의 분해 과정

단계	반지름	cosθ
2	[3 4 6 8]	
1	[5 10]	[3/5 3/5]
0	[$5\sqrt{5}$]	[$1/\sqrt{5}$]

재구성은 특징으로부터 원래 시퀀스를 구하는 과정으로 그림 2와 같이 복원된다. 예를 들어 첫 번째 특징 $5/\sqrt{5}$에 cosθ 값인 $1/\sqrt{5}$를 곱해 좌측의 5를 구하고, sinθ 값인 $2/\sqrt{5}$를 곱해 10을 구한다. 이런 식으로 좌측으로는 cosθ 값을 곱하고 우측으로는 sinθ 값을 곱해 내려가면 원래의 시퀀스 [3 4 6 8]을 구할 수 있다.

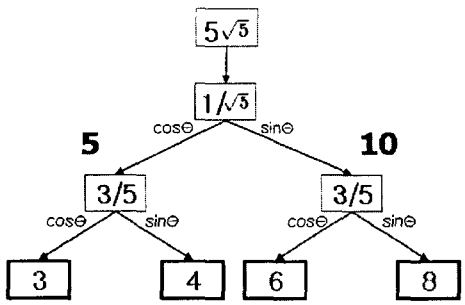

그림 2. Polar Wavelet의 재구성 과정

Harr wavelet에서 특징을 추출하기 위해서는 뒷부분에 0을 채워 전체 시퀀스 길이를 2^n으로 만들어야 한다. 그러나 Polar Wavelet은 각 단계에서 시퀀스의 마지막 홀수번째 데이타 값이 다음 단계로 그대로 유지되는 특성이 있으므로 2^n개까지 0을 채울 필요가 없다. 어떤 한 수와 0이 만드는 극좌표상의 반지름 값은 자신이 되기 때문이다. 그러므로 홀수번째 값을 그대로 다음단계의 반지름 값으로 사용하고 cosθ 값은 1로 하면 된다. 표 2를 보면 5번째 값인 10을 그대로 유지하고 conθ 값을 1로 하면서 특징을 구한다.

표 2. 길이가 2^n이 아닌 시퀀스의 분해과정

단계	반지름	cosθ
3	[3 4 6 8 <u>10</u>]	
2	[5 10 <u>10</u>]	[3/5 3/5 <u>1</u>]
1	[$5\sqrt{5}$ <u>10</u>]	[$1\sqrt{5}$ <u>1</u>]
0	[15]	[$\sqrt{5}/3$]

Polar Wavelet 기법은 반지름을 특징으로 추출하기 때문에 평균을 이용하는 Harr Wavelet보다 첫 단계에서의 후보 시퀀스가 줄어들게 된다는 것을 그림 3을 보면 알 수 있다. 그림 3은 길의 시퀀스가 [x, y]일 때, 두 Wavelet 기법의 첫 번째 특징 값인 평균, 반지름과 동일한 각 후보 시퀀스들의 예를 보여주고 있다. 각각의 시퀀스 내 변화의 정도나 패턴은 다르지만 전체적인 평균이 유사한 데이타의 경우, 예를 들어 연간 지역별 강수량이나 평균기온과 같은 시계열 데이타베이스에서는 후보 시퀀스 개수의 차이가 더 커질 것이다.

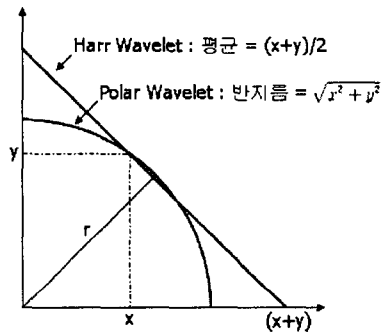

그림 3. Harr Wavelet과 Polar Wavelet 비교

3.2 착오 누락의 미 발생

특징 추출 기법에서 중요한 것은 차원을 감소시키면서도 유효한 시퀀스가 제외되는 착오누락이 발생하지 않는다는 것을 보장해야 한다는 것이다. 그러기 위해서는 감소된 차원인 특징 공간상의 거리가 실제 거리보다 작거나 같다는 것을 보장하기 위해 다음 조건을 만족시켜야 한다[1,2,5,6].

$$D_{feature}(A,B) \le D_{data}(A,B)$$

지면관계상 길이가 2인 경우에 대해 Polar Wavelet이 착오누락이 발생하지 않는다는 증명을 보이겠다.

두 시퀀스 $\vec{x}=(x_1, x_2)$, $\vec{y}=(y_1, y_2)$ 에서 Polar Wavelet에 의해 특징을 추출하면 다음과 같이 두 시퀀스 $\vec{a}$, $\vec{b}$ 를 얻을 수 있다.

$$P(\vec{x})=\vec{a}=(\sqrt{x_1^2+x_2^2}, \theta_1)=(r_1, \theta_1)$$

$$P(\vec{y})=\vec{b}=(\sqrt{y_1^2+y_2^2}, \theta_2)=(r_2, \theta_2)$$

각 시퀀스 사이의 유클리드 거리를 구하면 다음과 같이 된다.

$$D(\vec{x}, \vec{y})=(x_1-y_1)^2+(x_2-y_2)^2$$
$$=r_1^2+r_2^2-2r_1r_2cos(\theta_1-\theta_2) \quad ①$$

$$D(\vec{a}, \vec{b})=(r_1-r_2)^2+(\theta_1-\theta_2)^2$$
$$=r_1^2+r_2^2-2r_1r_2+(\theta_1-\theta_2)^2 \quad ②$$

특징 공간상의 거리가 실제 거리보다 커지지 않으면 즉, ① - ② ≥ 0 이면 착오누락이 발생하지 않는다는 것을 보장할 수 있다. $(\theta_1-\theta_2)=\alpha$ 라 하면 다음과 같은 식을 얻을 수 있다.

$$① - ② = r_1^2+r_2^2-2r_1r_2cos(\alpha)-r_1^2-r_2^2+2r_1r_2-\alpha^2$$
$$= 2r_1r_2(1-cos\alpha)-\alpha^2 \quad ③$$

시계열 데이타의 모든 데이타 값이 양수라고 가정하면 α의 범위는 $-\pi/2$와 $\pi/2$ 사이가 된다. ③식을 0보다 크게 하는 데이타 값을 구하기 위해 임의의 데이타를 x라 하면 $r=\sqrt{2x^2}$ 이 되며, x의 최소값을 구하면 $\pi/4$가 된다. 그러므로 시퀀스의 각 데이타 값이 $\pi/4$(≒0.785)이상일 경우 착오누락이 발생하지 않는다는 것을 알 수 있다.

3.3 질의 처리 방법

Polar Wavelet을 이용한 유사 검색 과정은 다음과 같다. 먼저 시계열 데이타베이스에 있는 모든 시퀀스에 Polar Wavelet을 적용하여 각 특징을 추출한 후, 이 특징을 이용하여 R-tree의 다차원 인덱스를 구성하는 전처리 과정을 수행한다. 주어진 질의 시퀀스에도 Polar Wavelet을 적용하여 특징을 추출한 후, 구성되어 있는 인덱스를 이용하여 유사한 후보 시퀀스를 검색한다. 각 후보 시퀀스에 대해서는 실제 유클리드 거리를 계산하는 후처리 과정을 통해 최종결과를 얻게 된다.

착오누락이 발생하지 않기 위해서는 데이타 값이 $\pi/4$ 보다 큰 시퀀스에 대해서만 적용할 수 있지만, $\pi/4$보다 큰 값이 되도록 질의 처리의 전처리 과정에서 데이타 값을 보정해 준다면 일반적인 모든 시퀀스에도 확장해서 적용할 수 있을 것이다.

5. 결 론

유사 검색 기법에서는 차원을 감소시키기 위해 데이타 시퀀스에서 k개의 특징을 추출하여 그 특징들을 k차원 공간으로 사상하는 특징 추출 기법을 사용한다. 효율적인 특징 추출 기법으로는 DFT, DWT, PAA 등이 제안되어 왔다.

본 논문에서는 L_2 norm인 극좌표를 사용하는 Polar Wavelet 기법을 제안하였고 특징 공간상의 거리와 실제 거리를 비교함으로써 시퀀스 내 데이타 값이 $\pi/4$ 이상일 경우 착오누락이 발생하지 않는다는 것을 증명하였다. 이 기법은 Harr Wavelet처럼 특징 추출을 위해 시퀀스의 길이를 2^n으로 만들지 않아도 되는 장점이 있으며 유사 검색시 후보 시퀀스의 개수를 줄임으로써 검색 성능을 향상시킬 수 있다.

참고 문헌

[1] Kin-pong Chan, and Ada Wai-chee Fu, Efficient Time Series Matching by Wavelet, in Proceedings of International Conference on Data Engineering (ICDE), pp. 126-133, 1999.

[2] Rakesh Agrawal, Christos Faloutsos, and Arun N. Swami, Efficient Similarity Search in Sequence Databases, in Proceedings of International Conference of Foundations of Data Organization (FODO), pp. 69-84, 1993.

[3] Byoung-Kee Yi, and Christos Faloutsos, Fast Time Sequence Indexing for Arbitrary Lp Norms, in Proceedings of International Conference on Very Large Data Bases (VLDB), pp. 383-394, 2000.

[4] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos, Fast Subsequence Matching in Time- Series Databases, in Proceedings of ACM SIGMOD Conference, pp. 419-429, 1994.

[5] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra, Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases, in Proceedings of Knowledge and Information Systems (KAIS), pp. 263-286, 2000.

[6] Ivan Popivanov, Renee J. Miller, Similarity Search over Time-Series Data Using Wavelets, in Proceedings of International Conference on Data Engineering (ICDE), pp. 212-221, 2002.