[1]시퀀스 데이타들 간의 관계성에 기반한 유사 검색 기법

강성구[O] 이석호
서울대학교 전기 컴퓨터 공학부
exodus[O]db.snu.ac.kr shlee@cse.snu.ac.kr

Association Based Similarity Search in Time Series DataBases

Seonggoo Kang[O] Sukho Lee
School of Electrical Engineering and Computer Science, Seoul National University

요 약

시퀀스 데이타는 크기를 가지는 일련의 값들로 이루어져 있어 일반적인 상품 데이타와는 달리 서로간의 관계성을 파악하기가 어려운 것으로 알려져 있다. 본 논문에서는 이러한 문제점을 해결하기 위하여 관계성을 보이는 시퀀스를 유사 시퀀스로 검색해 내는 기법을 제안한다. 이를 위해 유클리드 거리만으로 유사도가 결정되던 기존의 유사 검색을 변형하여 시퀀스의 상대적 위치와 형태를 고려한 시퀀스의 변화율을 척도로 사용하였으며 고차원이라는 문제를 해결하기 위하여 관계성을 수치로 표현하였다. 또한 본 논문에서는 기존의 하르 웨이블릿을 변형한 기하 웨이블릿을 이용하여 인덱스를 구성하였으며 보정 과정을 통해 기존의 유사 검색 기법으로도 문제가 변형될 수 있음을 보였다.

1. 서 론

존재 유무 만으로 관계 법칙을 유도하던 일반 상품 데이타[4]와는 달리 시퀀스 데이타는 크기를 가지는 일련의 값들로 이루어져 있어 관계성을 정의하기가 힘든 것으로 알려져 왔다. 이는 시퀀스가 가지는 고차원이라는 특징과 각 차원이 가지는 데이타 값의 비교 문제를 해결하지 않고서는 시퀀스 데이타들 간의 관계성을 정의하기가 힘들다는 것을 의미한다. 이 문제를 풀기 위한 하나의 방편으로 시퀀스를 다차원 상의 한 점으로 간주하여 점들간의 거리만으로 유사성을 판단하는 기존의 유사 검색[1, 2, 5, 6]을 생각해 볼 수 있다. 그러나 기존의 유사 검색에서는 질의 시퀀스와 값이 유사한 시퀀스만을 검색할 뿐 비슷한 형태 혹은 관계를 가지는 시퀀스를 검색하지 못하는 한계를 가진다.

시퀀스들 간의 관계는 시퀀스가 가지는 값 보다는 형태[2, 3]가 좌우한다. 시퀀스 데이타의 값이 서로 비슷하다고 해서 관계성이 성립하는 것이 아니라 서로 간의 영향력으로 인해 비슷한 형태를 보이는 시퀀스가 관계성이 깊다는 것이다. 예를 들어 한 주식 데이타의 변동이 다른 주식 데이타 값에 영향을 준다면 두 주식 데이타는 관계가 있다고 볼 수 있으나 주식 값이 비슷하다고 해서 두 주식이 관계성을 보이는 것은 아니다.

시퀀스 간의 형태가 관계성의 중요한 척도이긴 하나 시퀀스가 항상 같은 형태를 보이는 것은 아니다. 두 시퀀스의 관계가 긍정이냐, 부정이냐에 따라 시퀀스의 형태가 함께 증가하거나 감소할 수도 있고 그 반대의 경우가 발생할 수도 있다. 또한 시퀀스의 상대적 위치에 따라 시퀀스의 변화량이 다르다는 것에 주의해야 한다. 상대적으로 큰 데이타 값을 가지는 시퀀스 데이타의 변화량이 상대적으로 작은 데이타 값을 가지는 시퀀스보다는 클 수 밖에 없다. 따라서 시퀀스 간의 관계와 상대적 위치에 따라 시퀀스 간의 관계성을 재정립할 필요가 있다.

본 논문에서는 질의 시퀀스와 관계성이 높은 시퀀스를 데이타베이스내에서 검색해내는 기법을 제안한다. 이 기법은 시퀀스의 관계성이 시퀀스의 형태와 밀접하며 상대적 위치에 따라 변화 폭이 달라지는 것에 착안, 변화율이 비슷한 시퀀스를 관계성이 높은 시퀀스로 정의하였다. 또한 시퀀스가 가지는 고차원의 특징을 해결하기 위하여 관계성을 실수 값으로 표현하여 기존의 유사 검색으로의 응용이 가능하도록 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 기존의 유사 검색과 시퀀스의 형태에 관련된 연구들을 살펴보고 3절에서는 시퀀스 내에서 변화율을 특징으로 추출하는 기하 웨이블릿을 제안한다. 4절에는 이 문제가 보정 과정을 통해 기존의 유사 검색으로 변형될 수 있음을 보이고 5절에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련 연구

유사한 시퀀스를 찾아내는 문제는 DFT (Discrete Fourier Transform)를 사용하여 차원을 줄인 후 공간 접근 기법을 적용한 F-index[1]를 소개하면서 처음 시작되었다. ST-index[5]에서는 기존의 DFT로 차원을 감소하여 인덱스를 구성하였으며 이는 현재 유사 검색의 기본 틀이 되었다. 이후 차원을 감소하기 위한 방법으로 DWT (Discrete Wavelet Transform)[3], SVD(Singular Value Decomposition)[6] 기법이 사용되었다.

[1] 본 연구는 2005년도 두뇌한국21사업과, 정보통신부의 대학 IT연구센터(ITRC) 지원을 받아 수행되었습니다.

이후 시퀀스의 형태에 기반한 유사 검색이 연구되기 시작하였다. STB-index[3]에서는 시퀀스를 여러 개의 세그먼트로 나눈 후 상승이면 1, 하강이면 0으로 나타낸 후 이를 인덱싱 하였다. 그러나 이 기법은 유클리드 거리 입장에서 착오 누락이 발생한다는 단점이 있다. [2] 에서는 기존의 유사 검색에서는 값이 비슷한 시퀀스만을 찾아내므로 질의 시퀀스를 수직으로 이동시키면서 유사한 시퀀스를 찾아 내는 유사 검색 기법이 제안되었다. 그러나 이 기법은 시퀀스의 상대적 위치에 따라 변동 량이 달라질 수 있음을 고려하지 못한 문제점을 가지고 있다.

3. 변화율에 기반한 기하 웨이블릿 기법

이 절에서는 시퀀스 데이타 내에서 변화율이라는 특징을 추출하는 과정과 이를 인덱싱 하는 방법을 설명한다. 3.1절에서는 두 데이타로부터 얻어지는 기하평균으로부터 변화율을 얻을 수 있음을 보이고 3.2절에서는 이를 기반으로 하여 길이가 긴 시퀀스 데이타를 분해하는 기하 웨이블릿을 설명한다. 3.3절에서는 마지막으로 관계성에 기반한 유사 검색의 전체적인 질의 과정을 설명한다.

3.1 변화율과 기하 평균

어떠한 시퀀스에서 특징을 추출하기 위해서는 데이타의 변형을 필요로 한다. 예를 들어 길이가 2인 시퀀스 데이타 $\bar{x}=(x_1,x_2)$ 대하여 기존의 하르 웨이블릿에서는 이들의 산술 평균 $(x_1+x_2)/2$과 차이 $(x_1-x_2)/2$로 변형하여 특징을 추출하였다.

따라서 변화율이라는 특징을 추출하기 위해서는 또 다른 형태의 변형을 필요로 하게 된다. 그러나 여기서 주목할 것은 변화율이 비율 정보라는 것이다. 즉, 기존의 하르 웨이블릿이 덧셈에 대한 연산인 반면 변화율이라는 특징을 추출하기 위해서는 곱셈에 대한 연산이 필요하게 되는 것이다. 이를 위해 하르 웨이블릿에서 덧셈에 대한 산술평균을 구한 것처럼 아래와 그림 1과 같은 평균의 개념을 이용하여 곱셈 연산에 대한 평균을 구해 본다.

$$x \circ x = x_1 \circ x_2 \text{ , } (x \text{는 평균})$$

그림 1. ∘ 연산에 대한 평균의 계산

위의 식을 이용하면 x_1과 x_2의 곱셈에 대한 평균 $\sqrt{x_1x_2}$을 얻을 수 있게 된다. 하르 웨이블릿에서와 마찬가지로 x_1과 기하평균인 $\sqrt{x_1x_2}$의 비례를 살펴보면 $\sqrt{x_1/x_2}$의 값을 얻을 수 있게 되고 이 값은 x_1과 x_2 사이에서 발생하는 변화율의 제곱근이 됨을 쉽게 알 수 있다. 따라서 기존의 시퀀스 $\bar{x}=(x_1,x_2)$를 기하 평균과 변화율의 제곱근으로 변형할 수 있으며 변화율에 대한 비교가 가능해진다. 그림 2는 변화율을 추출하기 위한 기존 시퀀스 데이타의 변형을 식과 그림으로 나타낸 것이다.

3.2 기하 웨이블릿 기법

3.1절에서는 기하 평균을 이용하여 변화율을 얻어낼 수 있음을 증명하였다. 이번 절에서는 이러한 기하 평균을 이용하여 시퀀스 데이타를 변형하는 기하 웨이블릿 기법에 대하여 설명한다.

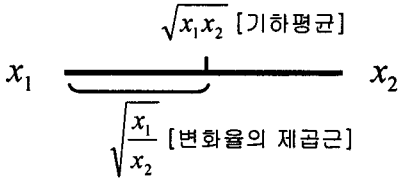

그림 2. 기하 평균을 이용한 변환 연산

다음의 표 3은 시퀀스 [2, 8, 16, 4]의 단계별 분해 과정을 보여준다.

표 3. 기하 웨이블릿의 분해 과정

단계(resol.)	기하 평균	변화율의 제곱근
3	[2 8 16 4]	
2	[4 8]	[0.5 2]
1	[$4\sqrt{2}$]	[$1/\sqrt{2}$]

기하 웨이블릿은 각 단계의 분해 과정에서 두 데이타끼리 짝을 지어 이들의 기하 평균과 변화율의 제곱근 값을 구하는 것으로 시작한다. 예를 들어 시퀀스 [2, 8, 16, 4]에서 데이타 2와 8을 하나의 짝으로 묶고 이들의 기하 평균 4와 변화율의 제곱근 0.5를 다음 해상도 값으로 정한다. 또한 16과 4의 기하 평균을 구하면 8이 되고 변화율의 제곱근은 2가 된다. 다음 해상도(resolution) 단계에서도 동일한 방법으로 4와 8의 기하 평균인 $4\sqrt{2}$와 변화율의 제곱근인 $1/\sqrt{2}$을 구한다. 이러한 기하 웨이블릿 분해 과정을 통해 얻어진 시퀀스 전체의 기하 평균과 각 단계에서 얻어진 변화율의 제곱근이 기하 웨이블릿의 특징으로 얻어지게 된다. 즉, [$4\sqrt{2}$,$1/\sqrt{2}$, 0.5. 2]가 변환된 웨이블릿 계수가 된다.

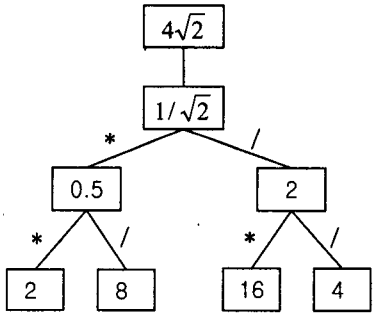

그림 4. 오차 나무를 이용한 데이타의 복원

이렇게 만들어진 기하 웨이블릿 계수를 이용하여 원본 데이타를 만들어내는 복원 과정은 그림 4의 오차 나무를 이용한다.

원본 시퀀스 $\vec{x}=(x_1,x_2)$는 3.1절의 기하 웨이블릿 변환을 통해 얻어진 기하 평균과 변화율의 제곱근을 이용하여 다음과 같이 얻어질 수 있다.

$$x_1=\sqrt{x_1x_2}\cdot\sqrt{x_1/x_2},\ x_2=\sqrt{x_1x_2}/\sqrt{x_1/x_2}$$

따라서 x_1을 구하기 위해서는 기하 평균과 변화율의 제곱근을 곱하고 x_2를 구하기 위해서는 기하 평균을 변화율의 제곱근으로 나누면 된다. 이를 $n-$차원 시퀀스로 확장하면, 오차 나무에서 왼쪽으로 갈 때는 변화율의 제곱근을 곱하고 오른쪽으로 갈 때는 반대로 나누어 주면 원본 데이타를 얻어낼 수 있다. 그림 4에서처럼 두 번째 데이타 8을 얻어낼 때에는 루트에서부터 왼쪽과 오른쪽 길을 한번씩 거쳐야 하므로 $4\sqrt{2}$에서 $1/\sqrt{2}$을 곱하고 다시 0.5로 나누어 주면 원본 데이타 8을 얻어낼 수 있게 된다.

3.3 질의 처리

이 절에서는 제안된 시퀀스 인덱싱 기법의 전체적인 과정을 설명한다. 먼저 질의가 수행되기 전 시계열 데이타베이스에 있는 모든 시퀀스에 대해 기하 웨이블릿 기법을 적용하여 특징을 추출한다. 그리고 이 추출된 특징들을 이용하여 R-tree 혹은 R*-tree과 같은 다차원 인덱스 구조에 저장한다. 그러나 기존의 인덱싱과 다른 한 가지는 기하 웨이블릿으로부터 얻어진 시퀀스 전체의 기하 평균, 즉 첫 번째 계수는 이용하지 않는다는 것이다. 이는 전체 기하 평균이 시퀀스의 상대적 위치 만을 판단하기 위한 정보일 뿐 시퀀스의 변화율을 통한 관계성과는 거리가 멀기 때문이다. 따라서 3.2절에서 얻어진 기하 웨이블릿 변환 $[4\sqrt{2},1/\sqrt{2},0.5,2]$에서 $4\sqrt{2}$를 제외한 $[1/\sqrt{2},0.5,2]$ 만으로 인덱스를 구성하도록 한다.

이후의 과정은 기존의 유사 검색 기법과 동일이다. 시계열 데이타베이스에 유사 검색 질의가 들어오게 되면 주어진 질의 시퀀스에도 동일한 기하 웨이블릿 기법을 적용하여 특징을 추출한다. 그리고 시퀀스 전체의 기하 평균이 제외된 다차원 인덱스 구조를 이용하여 질의 시퀀스와 유사한, 즉 질의 시퀀스와의 유클리드 거리가 주어진 허용치 ε보다 작은 후보 시퀀스들을 검색한다. 검색된 각각의 후보 시퀀스에 대해서는 데이타베이스의 시퀀스 데이타에 직접 접근하여 질의 시퀀스와의 실제 유클리드 거리를 계산한 후 거리가 ε이내인 시퀀스를 최종 결과로 출력 한다. 이 때 데이타 베이스에 존재하는 시퀀스는 원본 데이타이기에 변화율간의 유클리드 거리가 ε이내인 시퀀스를 정의하기가 힘들다. 따라서 다음 절의 보정 과정을 먼저 이야기 하고 이 문제를 언급하고자 한다.

4. 기존의 유사 검색으로의 변형

변화율을 이용한 문제는 보정 과정을 통해 기존의 유사 검색으로 변형할 수 있다. [2]에서는 질의 시퀀스를 수직으로 옮길 수 있도록 하기 위하여 하르 웨이블릿에서 얻어지는 첫 번째 계수, 즉 시퀀스 전체의 산술 평균을 제외함으로써 변화 량에 대한 비교 연산이 가능하도록 하였다. 변화율을 이용한 문제에서도 이와 마찬가지로 기하 웨이블릿에서 얻어진 웨이블릿 계수 중 시퀀스 전체의 기하 평균을 곱셈에 대한 항등원 1로 만들어 줌으로써 변화율 문제를 기존의 유클리드 거리 문제로 바꿀 수 있게 된다. 예를 들어 위의 3.2절의 예에서와 같이 얻어진 $[4\sqrt{2},1/\sqrt{2},0.5,2]$에서 시퀀스 전체의 기하 평균을 나타내는 $4\sqrt{2}$를 1로 만들어내면 모든 시퀀스가 동일한 점에서 시작된 변화율로 바뀌게 되어 절대 거리에 기반한 기존의 유사 검색 문제로 바뀔 수 있게 된다.

예를 들어 기존의 시퀀스 데이타 $\vec{x}=[2,8,16,4]$를 시퀀스의 기하 평균인 $4\sqrt{2}$로 나누면 보정된 데이타 $\vec{x}'=[1/2\sqrt{2},\sqrt{2},2\sqrt{2},1/\sqrt{2}]$ 을 얻게 되고 보정된 질의 시퀀스와 $\vec{x}'$의 유클리드 거리가 관계성과 관련된 수치와 같아진다. 즉, 3.3의 질의 처리 과정에서 얻어진 후보들을 데이타베이스에서 검색할 경우 각 시퀀스 데이타를 그들의 기하 평균으로 나눈 후 이들에 대한 유클리드 거리를 계산하면 실제 관계 있는 시퀀스를 얻어낼 수 있게 된다.

5. 결론 및 향후 과제

본 논문에서는 시퀀스의 관계성에 기반한 유사 검색 기법을 제안한다. 시퀀스 간의 관계는 시퀀스가 나타내는 형태와 밀접한 관련이 있으며 시퀀스 데이타의 상대적 위치에 따라 형태가 다르다는 것을 알 수 있었다. 따라서 변화량 보다는 시퀀스 들간의 변화율에 기반하여 유사성을 정의하였으며 이를 통해 관계성이 높은 시퀀스를 얻어낼 수 있었다. 또한 시퀀스 데이타를 기하 평균 값으로 나누어 줌으로써 기존의 유사 검색으로도 변형이 가능함을 보였다. 향후 연구 과제로는 시퀀스 데이타 간의 부정적인 영향에 기반한 검색과 여러 시퀀스에 의해 복합적으로 나타나는 관계성에 관한 연구를 진행할 것이다.

참고 문헌

[1] Rakesh Agrawal, Christos Faloutsos, Arun N. Swami, "Efficient Similarity Search In Sequence Databases", In Proceedings of FODO 1993: 69-84

[2] Kin-pong Chan, Ada Wai-chee Fu, "Efficient Time Series Matching by Wavelets", In Proceedings of ICDE 1999: 126-133

[3] Eamonn J. Keogh, Michael J. Pazzani, "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback" In Proceedings of KDD Conference 1998: 239-243

[4] Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499

[5] Christos Faloutsos, M. Ranganathan, Yannis. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases", In Proceedings of ACM SIGMOD Conference 1994: 419-429

[6] Flip Korn, H. V. Jagadish, Christos Faloutsos, "Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences", In Proceedings of ACM SIGMOD Conference 1997: 289-300