시계열 데이터베이스에서의 분해법을 이용한 유사 검색 기법

박 신 유[0] 문 봉 희
숙명여자대학교 컴퓨터과학과

Similarity Search in Time-Series Databases Using Decomposition Method

Shinyu Park[0] Bong-Hee Moon
Dept. of Computer Science, Sookmyung women's University

요 약

최근 몇 년간 시계열 데이터의 저장 및 분석에 대한 연구가 활발히 진행되고 있으며, 시계열 데이터베이스에서 유사패턴(similarity pattern)을 탐색하는 기법이 광범위한 응용분야에서 중요한 연구주제로 자리잡고 있다. 본 논문에서는 회귀분석방법을 바탕으로 한 분해 시계열 방법을 이용함으로써 기존의 유사성의 개념을 확장 시켰다. 즉, 시계열 데이터가 가지고 있는 패턴을 여러 성분으로 분해하여 각기 다른 저장 공간에 저장하고, 이를 이용하여 유사성을 탐색할 때에도 분리된 각 성분 중 특정 변동특성이 유사한 데이터를 추가적으로 요구되는 시간 없이 검색할 수 있다. 이는 전체 시계열 데이터를 이해하는데 뿐만 아니라 데이터를 예측하는 방법에도 유용하게 사용될 수 있다.

1. 서 론

일반적으로 시계열은 등간격으로 발생되어진 실수들의 연속된 집합을 뜻한다[6]. 시계열 데이터베이스에 저장된 시계열 데이터를 데이터 시퀀스(data sequence)라고 하며, 질의 시퀀스와 유사한 데이터 시퀀스를 검색하는 연산을 유사 시퀀스 매칭(similar sequence matching)이라고 한다[1,2,3,4]. 최근 몇 년간 이러한 시계열 데이터의 저장 및 분석에 대한 작업과 연구가 활발히 진행되고 있으며, 시계열 데이터간의 유사성 문제는 가장 관심을 끄는 문제 중의 하나이다[2]. 이러한 시계열 데이터베이스에서의 유사 탐색은 변환 패턴이 유사한 주식 목록들, 유사한 판매 패턴을 갖는 생산품들, 유사한 기온 패턴을 갖는 구간들, 유사한 음성 속도를 갖는 사람들을 발견하는데 적용된다.

시계열 데이터들은 여러 성분 - 추세성분, 계절성분, 순환성분, 불규칙성분 - 으로 구성되어 있다. 기존에 제안된 유사도 탐색 방법은 본래의 시계열 데이터를 가지고 주어진 데이터의 전체적인 성격과 유사한 시계열 데이터를 찾는데 주력하였다. 그러나 실제 주어진 데이터의 각 성분을 나누어 유사도를 탐색하면 다른 결과를 얻을 수 있을 것이다. 즉, 전체적인 성격을 통합하여 탐색하였을 경우 유사도가 떨어졌던 데이터가 특정 성분에서는 강한 유사도를 가질 수 있다

본 논문에서는 회귀분석 방법을 바탕으로 한 분해 시계열 방법을 이용하여 데이터의 변동 특성별로 분리된 변동특성을 각기 저장하는 방법을 사용한다. 하나의 시계열은 추세성분과 주기적 성분으로 나누어 표현되어 각기 다른 저장공간에 저장되며, 이를 이용하여 유사성을 탐색할 때에도 분리된 각 성분 중 특정 변동특성이 유사한 데이터를 찾을 수 있다. 각 성분에 대해 복합적으로 유사검색으로 할 경우 또한, 하나의 데이터의 각 성분이 분리되어 있으나 서로 다른 저장공간에 저장되어 병렬적으로 검색될 수 있으므로 추가적인 검색시간이 필요하지 않다.

2. 관련연구

시계열 데이터베이스에서의 유사검색기법은 크게 전체 매칭과 서브시퀀스 매칭으로 분류된다. 전자는 데이터베이스 내의 모든 시계열 데이터 시퀀스의 길이가 모두 같고 질의 시퀀스 또한 같은 길이를 갖는 것을 말하며, 후자의 경우는 질의 수열과 대상 수열의 길이가 다른 것으로 이 경우가 더 실질적이라고 할 수 있다. 유사 매칭 시퀀스에 대한 방법을 제시한 첫 연구인 [1]에서는 변형없는 전체 매칭에 대해 다루고 있으며, 유사 척도로써 유클리디안 거리(Euclidean distance)를 사용한다. 효율적인 검색을 위해 다

차원검색을 이용하며, Parseval의 정리에 의해 거짓 삭제(false dismissal)가 없다는 것을 보장한다. 즉, 각 데이터 시퀀스를 첫 번째의 DFT 계수인 하나의 점으로 대체하여 R*-트리를 이용한 F-index를 만드는 것을 제안한다. 여기서 각 시퀀스들은 N차원 공간의 포인트로 간주되었다. 만일 두 포인트의 유클리디안 거리가 임계치보다 작다면, 두 포인트에 해당하는 두 시퀀스는 유사하다고 간주하였다. 각 시퀀스가 하나의 포인트로 매핑 되므로, R*-트리를 색인구조로 사용하여 포인트들을 저장하였다. 시퀀스들은 각각 f개의 특성을 사용하는 f 차원의 포인트들로 표현된다. 특성추출에는 이산 퓨리에 변환(Discrete Fourier Transform, DFT)을 사용하였는데, DFT를 사용한 이유는 퓨리에 변환 이전과 마찬가지로 DFT가 유클리디안 거리를 유지하기 때문이다(Parseval의 정리). 변환 후 f 개의 데이터가 시퀀스를 나타내는데 사용된다.

[3]에서는 [1]의 연구를 확장하여 서브시퀀스 매칭을 다루고, 길이가 다른 시계열 데이터 시퀀스를 찾는 문제를 해결하였다. 즉, [1]을 일반화 시킨 것으로서, 서로 다른 길이의 데이터 수열들간의 유사성 탐색을 위해 ST-index를 제안한 방법이다. 여기서도 [1]에서와 마찬가지로 시퀀스의 특성을 추출하기 위해 DFT를 사용하였으나, 길이가 다른 시퀀스를 비교하기 위해 각 시퀀스를 가능한 모든 위치로부터 시작하는 w 크기의 슬라이딩 윈도우(sliding window)들로 나눈 후, 슬라이딩 윈도우들을 서로 비교하였다. 따라서 시퀀스의 특성을 하나의 포인트로 나타내는 대신, 특성공간의 포인트들의 트레일(trail)로 표현하게 된다. 이러한 포인트들을 색인에 올리기 위해, 트레일들은 서브트레일(subtrail)들로 나누고, 각 서브트레일들은 최소 경계 사각형(Minimal Bounding Rectangle, MBR)으로 표현한다. 이 질의 Q의 MBR과 교차하는 MBR을 갖는 시퀀스가 질의와 유사하다고 판단된다.

[2]는 잡음이나 스케일링(scaling) 또는 이행이 있는 시계열 데이터들간의 유사성 검색을 가능하게 해준다. [2]에서는 두 시퀀스가 유사한 서브시퀀스의 쌍을 충분히 가지고 있을 때, 두 시퀀스는 유사하다고 본다. 이 때, 서브시퀀스의 쌍들은 겹치지 않는 시간 순차적 쌍들(non-overlapping time-ordered pairs)이어야만 한다. 두 시퀀스를 비교하는 과정에서 시퀀스의 일부분은 아웃라이어(outlier)로 간주되어 제거될 수 있으며, 두 시퀀스의 시간 축에 반드시 정렬되어 있을 필요는 없다. 유사성 비교는 한 시퀀스로부터 임계치 이내에 다른 시퀀스가 있는지 검사함으로써 행해진다. 물론 아웃라이어는 무시된다.

[4]에서의 연구는 유사 개념의 확장이라는 점에서 앞서 살펴본 논문들과 같은 맥락에 있다고 볼 수 있는 것으로서, 이동 평균(moving average)의 방법이 있으며, DFT변환 외에 웨이블릿 변환(wavelet transform)을 이용하여 효과적인 탐색 기법과 시계열 데이터의 차원을 감소시키는 방법을 제안한 방법도 있다[5].

3. 분해 시계열 방법

시계열 데이터가 가지고 있는 패턴을 여러 성분의 부분 패턴으로 분해할 수 있다. 즉, 시계열 데이터를 추세요인(trend), 순환요인(cyclical) 그리고 불규칙요인(irregularity)으로 나누어 분석하는 방법을 분해 시계열 방법이라 한다[7]. 여기서 추세요인은 시계열 데이터의 장기적 형태를 나타내며, 순환요인은 경제 혹은 특정산업의 주기적 성쇠를 나타낸다. 계절요인은 요일마다 반복되거나, 일년 중의 각 월에 의한 변화, 사분기 시계열 데이터에서 각 분기에 의한 변화 등 고정된 주기에 의한 변동을 말하며 불규칙요인은 위 세 가지 요인으로 설명할 수 없는 오차로, 노이즈(noise)나 아웃라이어에 해당한다.

추세성분(Tt), 계절성분(St), 순환성분(Ct) 불규칙성분(It)에 대한 시계열 데이터(yt) 분해식은 다음과 같다.

$$yt = f(Tt, St, Ct, It)$$
$$yt = Tt + St + Ct + It$$

이는 시계열 데이터를 구성하는 요인들을 분석하는데 뿐만 아니라, 시계열 데이터 값을 예측하는 데도 유용하다. 그리고 시계열데이터가 가지고 있는 각 성분들을 분석할 수 있으므로 전체 시계열 데이터를 이해하는데 도움이 될 뿐만 아니라, 각 부분패턴을 외삽하여(extrapolating) 예측값들을 구한 후, 이 예측 값들을 더하여 데이터를 예측하는 방법에 사용된다.

본 논문에서는 앞서 언급한 시계열 데이터의 구성요소 중 불규칙요인은 노이즈로 생각해서 고려대상에 넣지 않고, 추세요인과 계절요인만을 고려한다.

본 논문에서 사용한 시계열데이터 분해 방법은 먼저 주어진 시계열데이터가 분산의 동질성을 가지도록 변환한다. 그 다음, 직교다항식을 사용하여 이 변환된 시계열 데이터의 추세요인을 추출한다. 마지막으로, 추세요인이 제거된 추세조정 시계열 데이터에 계절평균모형(또는 계절삼각함수모형)을 적용해서 계절요인을 추출한다. 이렇게 분해된 성분들은 빠른 검색을 위해 주어진 시계열 데이터를 주파수 영역 데이터로 변환해주는 빠른 퓨리에 변환을 하여 앞에 두 점씩을 추출하여 저장한다. 각 성분들은 각기 다른 저장공간에 저장되며 하나의 데이터에 대해 저장된 각 데이터베이스에 대한 인덱스는 같은 값을 갖게 된다. 이때, 하나의 데이터에 대한 분해성분이 서로 다른 저장공간에 저장되어 있으므로 병렬하게 검색할 수 있다. 검색시 R*-트리를 사용한다.

4. 실험

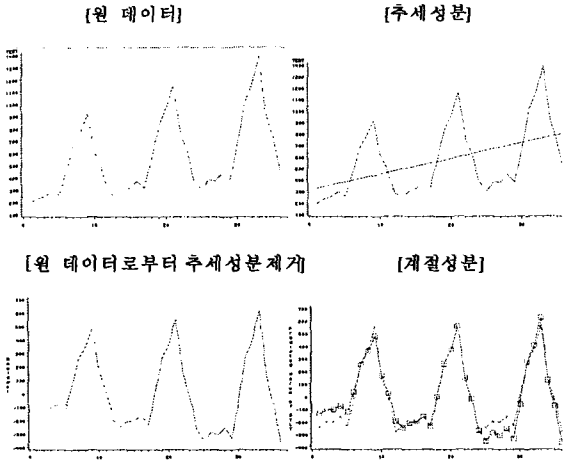
분해법을 적용하여 데이터를 추세성분과 계절성분으로 나누기 위해 SAS 6.12를 사용하였으며, 퓨리에 변환을 하여 데이터베이스를 구축하기 위해서 MATLAB5.3, Oracle7.3.3을 사용하였다. 마지막으로 주어진 시계열 데이터로부터 유사 시계열 데이터를 검색하기 위해 사용하는 R*-tree는 java를 이용하였다. 테스트한 데이터는 http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/에서 얻은 판매데이터이다.

실험 결과, 하나의 시계열전체를 두고 유사검색을 하는 기존 방법으로 실행하였을 경우 유사하다고 판단하지 않았

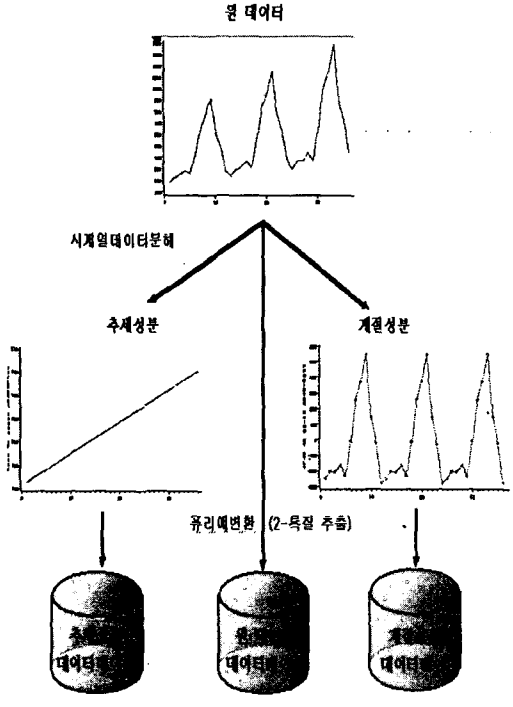
던 데이터에 대해 본 논문에서 제안한 분해법을 이용한 방법은 특정성분-추세성분이나 계절성분-에 있어서 유사하다고 판단하는 것을 살펴볼 수 있었다.

5. 예제

콜라 월별 판매 데이터를 가지고 추세성분과 계절성분으로 분해한 결과를 다음과 같다.



다음은 데이터베이스 구축과정을 그림으로 나타낸 것이다.



6. 결론 및 향후 연구 과제

시계열 데이터베이스에서의 유사 탐색 기법에 대한 기존의 연구에서는 하나의 시계열이 갖고 있는 복합된 성질이 모두 유사한 경우에 대해 탐색이 이루어졌다. 이에 반해, 본 논문에서는 시계열 데이터를 노이즈를 제거한 추세성분과 계절성분으로 나누어 각 성분에 대해 유사한 시계열 데이터를 탐색하게 된다. 따라서, 기존 연구에서는 찾을 수 없었던 특정 성분에서만 유사한 시계열들을 각기 분리하여 찾아볼 수 있으며, 이는 시계열데이터를 구성하는 요인들을 분석하는 데뿐만 아니라, 시계열 데이터 값을 예측하는 데도 유용하게 사용될 수 있다.

본 논문에서는 시계열 데이터가 가지고 있는 패턴 을 여러 성분의 부분패턴으로 분해하는 방법을 적용하여 유사도의 범위를 좀 더 확장 발전시킨 데 의의가 있다.

향후 연구 과제로는 추세성분보다 더 분리해내기 어려우나 원 데이터의 성격을 잘 반영할 수 있고 일상 데이터에서 좀 더 잘 관찰되는 주기가 분명치 않으나 반복되는 성격을 갖는 순환성분까지 고려하는 것에 대한 연구가 필요하다고 생각된다.

참고 문헌

[1] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient Similarity Search in Sequence Databases", In Proc. 4th Intl. Conf. On Foundations of Data Organization and Algorithms Conference, pp. 69-84, Chicago, Illinois, Oct. 1993.

[2] R. Agrawal, K -I. Lin, H.S. Sawhney, and K. Shim, "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time -Series Databases", In Proc. Int 'l C onf. On Very Large Databases, pp. 490 -501, Zurich, Switzerland, Sept. 1995.

[3] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time -Series Databases," In Proc. Int 'l conf. On Managament of Data, ACM SIGMOD, pp. 419 -429, Mi nneapolis, Minnesota, May 1994.

[4] D. Rafiei, and A. Mendelzon, "Similarity-Based Queries for Time Series Data ", In Proc. Int 'l Conf. On Management of Data, ACM SIGMOD, pp. 13 -25, Tucson, Arizona, May 1997.

[5] K.P. Chan and W.C. Fu. "Efficient Time Serie s Matching by Wavelets", In International Conference on Data Engineering", In Proc. the ACM PODS, Philadelphia PA, July 1999

[6] Byoung-kee Yi, H.V. Jagadish, and C. Faloutsos, "Efficient retrieval of similar time sequences under time warping ", In proc. In tl. Conf. On Data Engineering, Orlando, FL, Feburary 1998.

[7] 최병선, 단변량시계열분석1, 세경사, 1995