유사 시퀀스 매칭을 위한 하이브리드 저차원 변환

문 양 세[†] · 김 진 호[††]

요 약

유사 시퀀스 매칭에서는 고차원인 시퀀스를 저차원의 점으로 변환하기 위하여 저차원 변환을 사용한다. 그런데, 이러한 저차원 변환은 시계열 데이터의 종류에 따라 인덱싱 성능에 있어서 큰 차이를 나타낸다. 즉, 어떤 저차원 변환을 선택하느냐가 유사 시퀀스 매칭의 인덱싱 성능에 큰 영향을 주게 된다. 이 문제를 해결하기 위하여, 본 논문에서는 하나의 인덱스에서 두 개 이상의 저차원 변환을 통합하여 사용하는 하이브리드 접근법을 제안한다. 먼저, 하나의 시퀀스에 두 개 이상의 저차원 변환을 적용하는 하이브리드 저차원 변환의 개념을 제안하고, 변환된 시퀀스간의 거리를 계산하는 하이브리드 거리를 정의한다. 다음으로, 이러한 하이브리드 접근법 사용하면 유사 시퀀스 매칭을 정확하게 수행할 수 있음을 정형적으로 증명한다. 또한, 제안한 하이브리드 접근법을 사용하는 인덱스 구성 및 유사 시퀀스 매칭 알고리즘을 제시한다. 다양한 시계열 데이터에 대한 실험 결과, 제안한 하이브리드 접근법은 단일 저차원 변환을 사용하는 경우에 비해서 우수한 성능을 보이는 것으로 나타났다. 이 같은 결과를 볼 때, 제안한 하이브리드 접근법은 다양한 특성을 지닌 다양한 시계열 데이터에 두루 적용될 수 있는 우수한 방법이라 사료된다.

키워드 : 데이터 마이닝, 시계열 데이터베이스, 하이브리드 저차원 변환, 유사 시퀀스 매칭

Hybrid Lower-Dimensional Transformation for Similar Sequence Matching

Yang-Sae Moon[†] · Jinho Kim[††]

ABSTRACT

We generally use lower-dimensional transformations to convert high-dimensional sequences into low-dimensional points in similar sequence matching. These traditional transformations, however, show different characteristics in indexing performance by the type of time-series data. It means that the selection of lower-dimensional transformations makes a significant influence on the indexing performance in similar sequence matching. To solve this problem, in this paper we propose a hybrid approach that integrates multiple transformations and uses them in a single multidimensional index. We first propose a new notion of hybrid lower-dimensional transformation that exploits different lower-dimensional transformations for a sequence. We next define the hybrid distance to compute the distance between the transformed sequences. We then formally prove that the hybrid approach performs the similar sequence matching correctly. We also present the index building and the similar sequence matching algorithms that use the hybrid approach. Experimental results for various time-series data sets show that our hybrid approach outperforms the single transformation-based approach. These results indicate that the hybrid approach can be widely used for various time-series data with different characteristics.

Key Words : Data Mining, Time-Series Databases, Hybrid Lower-Dimensional Transformation, Similar Sequence Matching

1. 서 론

시계열 데이터(time-series data)란 각 시간별로 측정한 실수 값의 시퀀스로, 그 예로는 주식 데이터, 환율 데이터, 날씨 변동 데이터 등이 있다[1, 2, 3, 4]. 시계열 데이터베이스에 저장된 시계열 데이터를 데이터 시퀀스라 부르며, 사용자에 의해 주어진 시퀀스를 질의 시퀀스라 부른다. 그리고, 주어진 질의 시퀀스와 유사한 데이터 시퀀스를 검색하는 방법을 **유사 시퀀스 매칭**(*similar sequence matching*)이라 한다[2, 5]. 일반적으로, 유사 시퀀스 매칭에서는 길이 n인 두 시퀀스 $X(=\{x_0, x_1, \cdots, x_{n-1}\})$와 $Y(=\{y_0, y_1, \cdots, y_{n-1}\})$의 거리 함수 $D(X,Y)$로 유클리디안 거리($=L_2$)를 비롯하여, 맨하탄 거리($=L_1$), 최대 거리($=L_\infty$) 등의 L_p-거리($=\sqrt[p]{\sum_{i=0}^{n-1}|x_i-y_i|^p}$)를 주로 사용한다[1, 2, 3, 6, 7].

유사 시퀀스 매칭은 크게 전체 매칭(whole matching)과 서브시퀀스 매칭(subsequence matching)의 두 가지로 구분한다[2]. 전체 매칭은 질의 시퀀스와 유사한 데이터 시퀀스를 찾는 문제로서, 질의 시퀀스와 데이터 시퀀스의 길이가 동일한 특징을 갖는다[1]. 반면에, 서브시퀀스 매칭은 데이터 시퀀스에 포함된 서브시퀀스들 중에서 질의 시퀀스와 유사한 서브시퀀스를 찾는 문제로서, 사용자는 임의 길이의 시퀀스를 질의 시퀀스로 사용할 수 있다. 본 논문에서는 이러

※ 본 연구는 첨단정보기술연구센터를 통하여 과학기술부/한국과학재단의 지원을 받았음.
† 종신회원 : 강원대학교 IT특성화대학 컴퓨터과학전공 조교수
†† 정 회 원 : 강원대학교 IT특성화대학 컴퓨터과학전공 교수
논문접수 : 2007년 4월 26일, 심사완료 : 2007년 9월 5일

DOI: 10.3745/KIPSTD.2008.15-D.1.31

한 전체 매칭과 서브시퀀스 매칭 모두에 적용되는 **저차원 변환**(*lower-dimensional transformation*) 문제를 다룬다. 저차원 변환이란 고차원 공간의 점, 즉 고차원 시퀀스를 저차원 공간의 점으로 변환하는 기법으로, 많은 유사 시퀀스 매칭에서 사용되었다[1-10].

유사 시퀀스 매칭에서 저차원 변환을 사용하는 이유는 고차원인 시퀀스를 다차원 인덱스에 저장하기 위해서다[1, 2, 5, 8, 9, 10]. 즉, 다차원 인덱스의 고차원 문제[11]로 인하여, 고차원인 시퀀스를 R*-트리[12]와 같은 다차원 인덱스에 직접 저장하기 어렵기 때문에 저차원 변환을 사용한다. 여기에서 다차원 인덱스의 고차원 문제는 유사 시퀀스 매칭에서 "차원의 저주(curse of dimensionality)"라는 이름으로 잘 알려진 문제인데[11], 이는 다차원 인덱스에서는 차원이 높아질수록 검색 비용이 차원에 지수적으로 높아진다는 내용의 문제이다. 이러한 고차원 문제는 수십 혹은 수백 차원 이상의 고차원 데이터인 시계열 데이터를 직접 다차원 인덱스 저장하기 어렵게 하는 이유로 작용하였다. 이에 따라, 저차원 변환의 사용은 유사 시퀀스 매칭에서 인덱스의 사용을 가능하게 하였으며[1, 2], 이는 유사 시퀀스 매칭의 성능을 획기적으로 향상시키는 토대를 제공하였다. 이에 따라, 보다 우수한 인덱싱 성능을 얻기 위하여 다양한 저차원 변환에 대한 여러 연구가 진행되었다[2, 6, 8-10, 15, 17]. 그런데, 이러한 저차원 변환은 시계열 데이터의 특징에 따라서 인덱싱 성능에 큰 차이가 발생한다. 즉, 모든 시계열 데이터에 대해 최적의 인덱싱 성능을 발휘하는 특정한 저차원 변환은 없으며, 시계열 데이터의 특성에 따라 최적의 저차원 변환이 달라진다. 이러한 점에 착안하여 여러 개의 저차원 변환을 사용하여 여러 개의 인덱스를 구성하는 *E-Index*(Ensemble-Index) [10]가 제안되었다. E-Index는 여러 개의 인덱스를 의미하며, 각 저차원 변환 별로 별도의 인덱스를 구성한다. 그러나, E-Index는 미리 시계열 데이터의 특징을 파악해야 하는 단점과 여러 인덱스 관리에 따른 오버헤드로 인해 일반 사용자가 사용하기 어려운 문제점이 있다.

본 논문에서는 하나의 인덱스에서 두 개 이상의 저차원 변환을 통합하여 사용하는 하이브리드 접근법을 새로운 저차원 변환 기법으로 제안한다. 본 논문에서 "하이브리드"라는 용어를 사용한 이유는 제안한 접근법이 기존의 여러 저차원 변환을 통합하여 사용하기 때문이다. 즉, 두 개 이상의 기존 저차원 변환을 통합하는 새로운 저차원 변환이라는 개념에서 하이브리드 접근법이라 명명하였다. 저자들이 아는 한 유사 시퀀스 매칭의 저차원 변환을 위하여 하이브리드 접근법을 사용한 것은 본 논문이 처음이다. 제안한 하이브리드 접근법은 대다수의 저차원 변환이 대부분의 에너지를 소수 개의 특성(차원)에 집중한다[1, 6]는 관찰에서 출발한다. 즉, 여러 저차원 변환을 사용하되, 각 저차원 변환 별로 적은 수의 특성만을 추출한 후 이를 통합하여 인덱싱을 수행하는 방법이다. 이를 위해, 본 논문에서는 우선 **하이브리드 저차원 변환**(*hybrid lower-dimensional transformation*)의 개념을 제안한다(혹은 간략히 **하이브리드 변환**이라 한다). 하이브리드 저차원 변환은 두 개 이상의 저차원 변환을 사용하여 하나의 시퀀스에서 서로 다른 특징을 지닌 여러 특성들을 추출하는 저차원 변환 방법이다. 그리고, 이러한 하이브리드 변환을 유사 시퀀스 매칭에 사용하기 위해서, 하이브리드 변환된 두 시퀀스간의 거리를 계산하기 위한 **하이브리드 거리**(*hybrid distance*)를 정의한다. 다음으로, 이러한 하이브리드 변환과 하이브리드 거리를 사용하면 유사 시퀀스 매칭을 정확하게 수행할 수 있음을 정리로서 제시하고 증명한다. 또한, 제안한 하이브리드 변환과 하이브리드 거리를 사용한 인덱스 구성 알고리즘과 유사 시퀀스 매칭 알고리즘을 제시한다. 다양한 시계열 데이터에 대한 실험 결과, 제안한 하이브리드 접근법은 단일 저차원 변환을 사용하는 경우에 비해 우수한 성능을 보이는 것으로 나타났다. 특히, 하이브리드 저차원 변환은 시계열 데이터의 종류나 선택율 범위에 관계없이 단일 저차원 변환을 사용하는 경우보다 우수한 성능을 보이는 것으로 나타났다. 이 같은 결과를 종합하면, 데이터나 선택율에 따라 어떤 저차원 변환을 선택할 고민이 없이, 여러 저차원 변환을 통합하는 하이브리드 변환을 사용하면 유사 시퀀스 매칭에 있어서 보다 향상된 성능을 얻을 수 있음 의미한다. 이 같은 결과를 볼 때, 제안한 하이브리드 접근법은 다양한 특성을 지닌 다양한 시계열 데이터에 두루 적용될 수 있는 우수한 방법이라 사료된다.

본 논문의 구성은 다음과 같다. 제2장에서는 유사 시퀀스 매칭과 저차원 변환의 관련 연구를 설명한다. 제3장에서는 하이브리드 저차원 변환과 하이브리드 거리의 개념을 제안한다. 제4장에서는 제안한 하이브리드 접근법을 사용한 다차원 인덱스 구성 및 유사 시퀀스 매칭 알고리즘을 설명한다. 제5장에서는 실험을 통해 제안한 하이브리드 접근법의 우수성을 보인다. 마지막으로, 제6장에서 결론을 맺는다.

2. 관련 연구

본 장에서는 유사 시퀀스 매칭과 저차원 변환 관련 기존 연구를 설명한다. 저차원 변환을 포함하여 본 논문에서 사용하는 주요 표기와 이에 대한 정의 및 의미는 <표 1>과 같다.

〈표 1〉 주요 표기법

기호	정의/의미
F_i	i-번째 저차원 변환
$F_i(S)$	고차원 시퀀스 S가 저차원 변환 F_i에 의해 변환된 저차원 점
$F_i(S)F_j(S)$	저차원 변환된 두 점 $F_i(S)$와 $F_j(S)$의 연결(concatenation)
D(S,Q)	두 시퀀스 S와 Q에 대한 유클리디안 거리 함수(혹은 L_p-거리 함수)

서론에서 설명한 바와 같이 유사 시퀀스 매칭은 전체 매칭과 서브시퀀스 매칭으로 나눌 수 있다. 먼저, Agrawal 등[1]의 전체 매칭을 인덱스 구성 알고리즘과 유사 시퀀스 매칭 알고리즘으로 구분하여 설명한다. 인덱스 구성 알고리즘에서는 길이 n인 데이터 시퀀스에서 $f(\ll n)$개의 특성을 추출하여 f-차원 공간의 점으로 저차원 변환[1,2]한 후, 이를 f-차원의 다차원 인덱스에 저장한다. 이렇게 저차원 변환을 수행하는 이유는 다차원 인덱스의 고차원 문제[11]로 인하여, 고차원인 시퀀스를 다차원 인덱스에 직접 저장하기 어렵기 때문이다. 서론에서 언급한 바와 같이, 고차원 문제는 차원의 저주로 알려진 문제로서 다차원 인덱스의 검색 비용은 저장할 데이터의 차원에 지수적으로 증가한다는 내용이다. 이에 따라, 수십 혹은 수백 차원의 시계열 데이터에 대해서 다차원 인덱스를 사용하기 위해서는 저차원 변환의 사용이 불가피하다. 이와 같은 저차원 변환을 위해 사용하는 함수를 특성 추출 함수라 한다[2, 3, 5, 6]. 다음으로, 유사 시퀀스 매칭 알고리즘에서는 질의 시퀀스를 데이터 시퀀스와 동일한 방법으로 f-차원 점으로 변환하고, 변환한 점과 허용치 ε을 사용하여 범위 질의를 구성한다. (혹은 주어진 k를 사용하여 k-NN(nearest neighbor) 질의를 구성한다.) 그리고, 구성한 질의로 다차원 인덱스를 검색하여, 후보 시퀀스들을 구한다. 마지막으로, 각 후보 시퀀스 대해서는 데이터베이스에 저장된 실제 데이터 시퀀스를 액세스하고 질의 시퀀스와의 거리를 조사하여 실제 유사 시퀀스만을 판별하는 후처리 과정[1]을 수행한다.

다음으로, Faloutsos 등[2]은 전체 매칭을 일반화하여 서브시퀀스 매칭을 처음 소개하고, 이의 해결책(간략히 *FRM*이라 한다)을 제시하였다. FRM에서는 데이터 시퀀스를 슬라이딩 윈도우로 나누고 질의 시퀀스를 디스조인트 윈도우로 나누는 방법을 사용하며, 전체 매칭과 마찬가지로 인덱스 구성 알고리즘과 서브시퀀스 매칭 알고리즘으로 구성된다. 먼저, 인덱스 구성 알고리즘에서는 데이터 시퀀스를 나눈 슬라이딩 윈도우로 f-차원의 점으로 변환하여 다차원 인덱스에 저장한다. 다음으로, 질의 시퀀스를 나눈 디스조인트 윈도우를 f-차원의 점으로 변환하고, 이 점을 기준으로 범위 질의 혹은 k-NN 질의를 구성한다. 그리고, 다차원 인덱스를 검색하여 후보 시퀀스들을 찾아내고, 후처리 과정을 통하여 실제 유사 서브시퀀스만을 찾는다.

앞서의 전체 매칭과 서브시퀀스 매칭이 소개된 이후로 다양한 매칭 알고리즘들이 제시되었다. 먼저, 서브시퀀스 매칭 분야에서는 DualMatch[8]와 GeneralMatch[4]가 제안되었는데, 이들 방법은 윈도우 구성법을 달리하여 FRM의 성능을 개선하였다. 다음으로, 저차원 변환의 종류를 달리한 연구로서 Chan 등의 연구[6], Keogh 등의 연구[8-10]가 수행되었다. 또한, 정규화[7], 이동평균[14] 등의 전처리 기법을 지원하는 연구가 수행되었으며, 타임 워핑 거리[4], Lp-거리[16] 등의 다양한 거리를 사용하는 연구[4, 16]가 수행되기도 하였다.

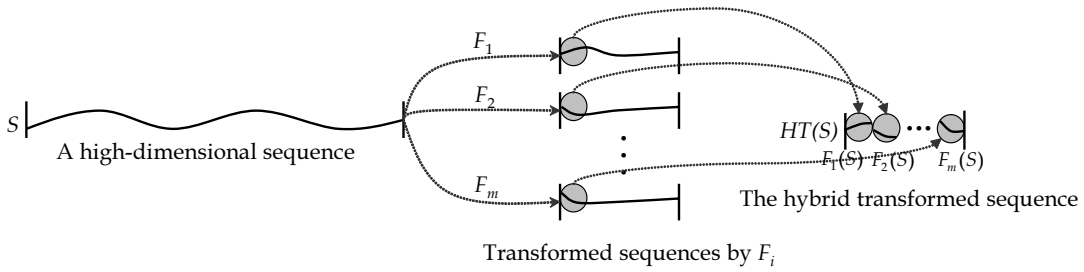
지금까지 설명한 대부분의 유사 시퀀스 매칭 방법은 고차원인 시퀀스를 다차원 인덱스에 저장하기 위하여 저차원 변환을 사용한다[1, 2, 6, 13, 14, 15]. 이러한 저차원 변환은 유사 시퀀스 매칭에서 검색 속도의 향상을 위해서 다차원 인덱스를 사용하는데, 그 사용 이유는 다차원 인덱스의 고차원 문제를 피하고 인덱스의 저장 공간을 줄이기 위함이다. 이와 같이 저차원 변환은 유사 시퀀스 매칭에서의 인덱스 사용을 가능하게 하였으며, 이러한 인덱스 사용의 효과를 극대화하기 위하여 보다 인덱싱 성능이 우수한 저차원 변환에 대한 많은 연구가 진행되었다[2, 6, 8-10, 15, 17]. 본 논문의 제3장 및 제4장에서 제안하는 하이브리드 접근법은 이러한 기존의 저차원 변환들의 장점을 취합하여, 보다 인덱싱 성능을 높이자는데 그 목적이 있다. 이와 같이 인덱싱 성능을 높일 수 있다면 궁극적으로 유사 시퀀스 매칭의 전체 성능을 크게 향상 시킬 수 있기 때문이다.

고차원 시퀀스의 저차원 변환 방법으로는 DFT(Discrete Fourier Transform), DWT(Discrete Wavelet Transform), PAA(Piecewise Aggregate Approximation) 등 여러 가지 변환이 사용되었다. 우선, DFT는 참고문헌 [1, 2, 5, 7, 13, 14, 15] 등 많은 연구에서 가장 널리 사용되었다. 다음으로, DWT 역시 참고문헌 [3, 6] 등에서, PAA는 참고문헌 [8, 16] 등에서 유사 시퀀스 매칭의 저차원 변환 방법으로 사용되었다. 이외에도, DCT(Discrete Cosine Transform)[17], SVD(Singular Value Decomposition)[9] 등 여러 가지 저차원 변환 방법이 제시되었다. 그런데, 이들 저차원 변환 중에서 특정 변환 하나를 최적으로 지정할 수는 없으며, (시계열) 데이터의 특성이나 종류에 따라 가장 적합한 변환이 달라질 수 있다[10].

시계열 데이터의 특징에 따른 최적의 저차원 변환을 사용하기 위하여 Keogh 등[10]은 E-Index를 제안하였다. E-Index는 특정한 하나의 인덱스가 아닌 여러 개의 인덱스를 의미하며, 각 저차원 변환 별로 별도의 인덱스를 구성한다. E-Index에서는 실험을 통해 시계열 데이터의 각 부분에 가장 적합한 저차원 변환을 선택하고, 선택한 변환에 따라 시계열 데이터를 여러 인덱스에 나누어 저장한다. 그러나, 이러한 E-Index는 실용적으로 사용하기 어려운 문제점이 있다. 첫째, 시계열 데이터의 종류와 특징은 무한할 뿐 아니라 지속적으로 갱신 혹은 추가될 수 있기 때문에, 이를 미리 예측하여 각 부분에 대한 최적의 인덱스를 구성하기는 매우 어렵다. 둘째, 시계열 데이터의 각 부분에 대해서 최적의 저차원 변환을 선택하고, 이를 바탕으로 인덱스를 구성하는 작업을 일반적인 사용자에게 적용하기는 매우 어렵다. 셋째, E-Index는 내부적으로 여러 인덱스가 구성되므로, 여러 인덱스 관리에 따른 오버헤드가 뒤따른다. 따라서, 본 논문에서는 하나의 인덱스에서 두 개 이상의 저차원 변환을 통합하여 사용하는 하이브리드 접근법을 제안한다. 제안하는 하이브리드 접근법은 각 시계열 데이터에 대해서 두 개 이상의 인덱스를 모두 사용하므로, 상기의 문제점이 발생하지 않는다. 즉, 각 부분에 적합한 저차원 변환을 분석/예측할 필요가 없으며, 하나의 인덱스만을 구축하므로 인덱스 오버헤드도 발생하지 않는다.

3. 하이브리드 저차원 변환

본 논문에서는 DFT, DWT, PAA 등의 많은 저차원 변환이 대부분의 에너지를 소수의 특성에만 집중한다는 점에 착



(그림 1) 하이브리드 저차원 변환의 개념.

안하여 새로운 저차원 변환을 제안한다. 즉, 여러 저차원 변환의 결과인 여러 특성들을 하나로 통합하는 새로운 저차원 변환을 정의한다.

정의 1: 주어진 m개의 저차원 변환 F_1, F_2, …, F_m에 대해서, 고차원 시퀀스 S를 **하이브리드 저차원 변환**(간략히 **하이브리드 변환**)한 시퀀스 $HT(S)$는 다음 식 (1)과 같이 정의한다.

$$HT(S) = F_1(S)\ F_2(S)\ \cdots\ F_m(S) \tag{1}$$

□

(그림 1)은 정의 1의 하이브리드 저차원 변환을 도식적으로 나타낸 것이다. 그림을 보면, m개의 저차원 변환인 F_1, F_2, ..., F_m을 하이브리드 변환에 사용함을 알 수 있다. 이와 같이 두 개 이상의 저차원 변환을 통합하여 하나의 새로운 저차원 변환으로 사용하기 때문에, 본 논문에서는 "하이브리드"라는 용어를 사용하였다. 즉, 여러 개의 기존 저차원 변환을 통합한다는 의미에서 하이브리드 저차원 변환이라 정의하였다. 하이브리드 저차원 변환이 수행되는 과정을 설명하면 다음과 같다. 먼저, 주어진 시퀀스 S를 각 F_i를 사용하여 저차원 변환된 점 $F_i(S)$를 구성한다. 즉, F_1, F_2, ..., F_m을 적용하여 m개의 저차원 변환된 점인 $F_1(S)$, $F_2(S)$, ..., $F_m(S)$을 구성하는 것이다. 그런 다음, 이들 m개의 점들을 연결하여 하나의 저차원 점인 $F_1(S)F_2(S)...F_m(S)$을 구성한다. 마지막으로, 이렇게 변환된 저차원 점 $F_1(S)F_2(S)...F_m(S)$를 다차원 인덱스의 인덱싱이나 범위 질의(혹은 k-NN 질의) 구성에 사용한다. 결국, (그림 1)은 주어진 시퀀스 S를 각 F_i를 통해 저차원 변환한 후, 각 저차원 변환된 시퀀스에서 에너지가 집중된 소수의 특성들만을 선택하여 하이브리드 변환된 시퀀스를 구성하는 방법을 나타낸다. 이와 같이 하이브리드 변환을 사용하는 이유는 각각의 저차원 변환이 서로 다른 특징을 가지므로[10], 하나의 인덱스에서 이들 여러 저차원 변환에 의해 추출한 특성들을 통합하여 사용하기 위해서이다. 즉, 하나의 시퀀스에 대해 여러 저차원 변환을 통해 다양한 특성들을 추출하여 통합하여 사용함으로써, 여러 저차원 변환의 장점을 고루 발휘하자는데 그 목적이 있다.

그런데, 하이브리드 저차원 변환을 유사 시퀀스 매칭에 사용하기 위해서는 변환된 두 시퀀스를 대상으로 하는 새로운 거리 계산 방법이 필요하다. 그 이유는 하이브리드 변환에 있어서 기존의 유클리디안 거리(혹은 L_p-거리) 계산을 그대로 사용할 경우, Parseval의 정리[1, 2][1]가 더 이상 만족하지 않기 때문이다. 즉, 주어진 두 시퀀스 S와 Q에 대해서 다음 공식 (2)가 더 이상 성립하지 않기 때문이다.

$$D(S,Q) \le \varepsilon \ \Rightarrow\ D(HT(S), HT(Q)) \le \varepsilon \tag{2}$$

저차원 변환을 유사 시퀀스 매칭에 사용하기 위해서는 공식 (2)의 관계가 반드시 만족되어야 한다[2, 3]. 따라서, 하이브리드 저차원 변환을 유사 시퀀스 매칭에 사용하기 위해서는 변환된 두 시퀀스 간의 새로운 거리 함수가 필요하다. 본 논문에서는 하이브리드 저차원 변환을 유사 시퀀스 매칭에 사용하기 위하여 다음과 같이 새로운 거리 함수를 정의한다.

정의 2: 두 시퀀스 S와 Q가 저차원 변환 F_1, F_2, …, F_m에 의해 각각 $HT(S)$와 $HT(Q)$로 하이브리드 저차원 변환되었다 할 때, 하이브리드 변환된 두 시퀀스 $HT(S)$와 $HT(Q)$간의 **하이브리드 거리** $HD(HT(S), HT(Q))$는 다음 공식 (3)과 같이 정의한다.

$$HD(HT(S), HT(Q)) \equiv \max_{1 \le i \le m} \{D(F_i(S), F_i(Q))\} \tag{3}$$

□

정의 2의 하이브리드 거리의 의미는 하이브리드 저차원 변환을 사용할 때, 여러 저차원 변환 중에서 변환된 두 시퀀스 간의 거리를 최대로 하는 저차원 변환을 선택함을 의미한다. 즉, 하이브리드 변환된 두 시퀀스를 비교할 때 가장 큰 거리 값을 인덱싱에 사용하자는데 그 목적이 있다. 이와 같이 가장 큰 거리 값을 하이브리드 거리로 사용하게 되면, 궁극적으로 유사 시퀀스 매칭에 있어서 보다 높은 인덱싱 성능을 나타낼 수 있기 때문이다[2, 3].

다음 정리 1은 하이브리드 거리를 사용할 경우 하이브리드 저차원 변환이 Parseval의 정리(정확히는 공식 (2))를 만족함을 나타낸다.

1) Parseval의 정리는 두 시퀀스와 저차원 변환이 주어졌을 때, 저차원 변환된 두 시퀀스 간의 거리가 저차원 변환되기 이전의 두 시퀀스 간의 거리보다 작거나 같아야 한다는 성질이다. 즉, 두 시퀀스가 S와 Q로 주어지고, 저차원 변환이 F로 주어졌을 때, $D(F(S), F(Q))$는 $D(S,Q)$보다 작거나 같아야 한다는 성질이다. 이러한 Parseval의 정리가 만족하여야만 유사 시퀀스 매칭에서 해당 저차원 변환을 인덱싱에 활용할 수 있다[1, 2, 5, 6].

정리 1: 저차원 변환 F_1, F_2, …, F_m이 각각 Parseval의 정리를 만족한다 할 때(공식 (2)를 만족한다 할 때), 두 시퀀스 S와 Q 사이의 거리 $D(S,Q)$와 F_1, F_2, …, F_m에 의해 하이브리드 저차원 변환된 두 시퀀스 $HT(S)$와 $HT(Q)$의 하이브리드 거리 $HD(HT(S),HT(Q))$ 사이에는 다음 공식 (4)의 관계가 성립한다.

$$D(S,Q) \le \varepsilon \Rightarrow HD(HT(S),HT(Q)) \le \varepsilon \quad (4)$$

증명: 가정에 의해 각 저차원 변환 F_i는 Parseval의 정리를 만족하므로, 공식 (2)에 의해 모든 F_i는 다음 공식 (5)의 관계를 만족한다.

$$D(S,Q) \le \varepsilon \Rightarrow D(F_i(S),F_i(Q)) \le \varepsilon \quad (5)$$

그런데, 모든 F_i에 대해서 공식 (5)가 성립해야 하므로, 다음의 공식 (6) 또한 성립한다.

$$D(S,Q) \le \varepsilon \Rightarrow \max_{1 \le i \le m} \{D(F_i(S),F_i(Q))\} \le \varepsilon \quad (6)$$

여기에서 $\max_{1 \le i \le m} \{D(F_i(S),F_i(Q))\}$는 정의 2에 의해 $HD(HT(Q), HT(S))$로 정의되므로, 결국 공식 (4)가 성립함을 알 수 있다. □

정리 1의 공식 (5)는 유사 시퀀스 매칭에서 사용하는 하한 정리(lower-bounding theorem) 조건[4]을 나타낸다. 이러한 하한 정리의 의미는 필요 조건이 만족하면 충분 조건도 만족한다는 것으로, 공식 (5)에서 오른편 식인 필요 조건이 만족하면, 공식 (5)에서 왼편 식인 충분 조건은 반드시 만족한다는 의미이다. 이는 필요 조건이 만족하는 시퀀스들을 찾아 후보로 삼으면 착오기각이 발생하지 않음(충분 조건)을 보장한다는 의미로서, 많은 유사 시퀀스 매칭[1-3, 5-8, 13-16]에서는 이와 유사한 하한 정리를 정확성의 근거로 제시하였다. 결국, 정리 1의 의미는 하이브리드 저차원 변환을 위한 거리 함수로서 하이브리드 거리를 사용한다면 유사 시퀀스 매칭을 정확히 수행할 수 있음을 나타낸다. 이에 따라, 다음 제4절에서는 하이브리드 저차원 변환과 하이브리드 거리를 사용한 인덱스 구성과 유사 시퀀스 매칭 알고리즘을 설명한다.

4. 인덱스 구성 및 유사 시퀀스 매칭 알고리즘

본 절에서는 하이브리드 저차원 변환과 하이브리드 거리를 사용한 유사 시퀀스 매칭을 설명한다. 먼저, 제4.1절에서는 하이브리드 저차원 변환을 사용한 다차원 인덱스 구성과 알고리즘을 제시한다. 다음으로, 제4.2절에서는 하이브리드 거리를 사용한 유사 시퀀스 매칭과 알고리즘을 제시한다.

4.1 인덱스 구성 알고리즘

유사 시퀀스 매칭을 위해서는 먼저 저차원 변환을 사용하여 다차원 인덱스를 구성해야 한다. 하이브리드 저차원 변환을 사용하여 인덱스를 구성하는 방법은 기존 저차원 변환을 사용하는 방법[1, 2, 5, 6, 13, 16]과 동일하다. 즉, 고차원인 데이터 (서브)시퀀스를 하이브리드 변환을 사용하여 저차원의 점(혹은 MBR)으로 변환하고, 이들 저차원 변환된 점(혹은 MBR)을 다차원 인덱스에 저장하는 것이다. 이와 같이 구성한 다차원 인덱스는 추후 유사 시퀀스 매칭에서 후보(candidate) 데이터 시퀀스를 구하는데 사용된다.

```
Algorithm BuildIndex(𝕊, F_1,…,F_m)
// 𝕊 is a set of data sequences to be indexed.
// F_i is the i-th lower-dimensional transformation of the hybrid transformation.
1  for each data sequence S ∈ 𝕊 do
2      for each lower-dimensional transformation F_i do
3          Obtain the low-dimensional point F_i(S) using F_i;
4      end for
5      Obtain the hybrid transformed point HT(S) by concatenating F_1,…,F_m;
6      Construct a record <HT(S), ptr(S)>, where ptr(S) is the pointer to S;
7      Insert the record <HT(S), ptr(S)> into the index;
8  end-for
```

(그림 2) 하이브리드 저차원 변환을 사용한 인덱스 구성 알고리즘

(그림 2)은 하이브리드 저차원 변환을 사용하여 다차원 인덱스를 구성하는 알고리즘을 나타낸다. (그림 2)의 알고리즘은 전체 매칭을 위한 인덱스 구성으로서, 참고문헌 [2, 3, 5, 13]등에 의해 서브시퀀스 매칭을 위한 인덱스 구성으로 쉽게 확장할 수 있다. 그림을 보면, 알고리즘의 입력으로는 디스크에 저장된 데이터 시퀀스들의 집합과 하이브리드 변환에 사용할 저차원 변환들이 주어진다. 알고리즘의 라인 2~4에서는 주어진 저차원 변환들인 F_1, F_2…, F_m을 사용하여 데이터 시퀀스를 $F_1(S)$, $F_2(S)$…, $F_m(S)$의 저차원 점들로 변환한다. 그런 다음, 라인 5에서는 이들 저차원 점들을 연결하여 하나의 저차원 변환된 점인 $HT(S)$를 구성한다. 결국 이러한 라인 2~5의 과정이 각 고차원 시퀀스에 대해서 하이브리드 변환을 수행하는 과정이다. 라인 6과 7은 하이브리드 변환된 점을 인덱싱하는 과정이다. 먼저, 라인 6에서는 변환된 점과 해당 시퀀스에 대한 포인터를 사용하여 다차원 인덱스에 저장할 레코드를 구성한다. 다음으로, 라인 7에서 구성한 레코드를 다차원 인덱스에 저장한다. 이러한 하이브리드 변환 및 인덱스 저장 과정을 각 시퀀스에 대해 반복하여 다차원 인덱스을 구성할 수 있다(라인 1~8).

4.2 유사 시퀀스 매칭 알고리즘

하이브리드 저차원 변환을 사용한 유사 시퀀스 매칭을 수행하기 위해서는 다차원 인덱스에 대한 인덱스 검색 방법이 달라져야 한다. 기존의 단일 저차원 변환을 사용하는 방법에서는 인덱스 검색 과정에서 변환된 두 시퀀스 간의 거리 함수로서 통상의 유클리디안 거리(혹은 L_p-거리)를 사용한다. 반면에, 하이브리드 저차원 변환을 사용하기 위해서는 정의 2의 하이브리드 거리(혹은 하이브리드 L_p-거리)를 사용해야 한다. 이와 같은 변화가 필요한 이유는 단일 저차원 변환이 일반적인 L_p-거리에 대해서 Parseval의 정리(즉, 공

```
Algorithm WholeMatching(Q,ε,F1,…,Fm)
//  Q is a query sequence given by a user.
//  ε is the user-specified tolerance.
//  Fi is the i-th lower-dimensional transformation of the hybrid transformation.
1   for each lower-dimensional transformation Fi do
2       Obtain the low-dimensional point Fi(Q) using Fi;
3   end for
4   Obtain the hybrid transformed point HT(S) by concatenating F1,…,Fm;
5   Construct a range query using HT(Q) and the tolerance ε;
6   Obtain the candidate sequences by searching the index;  // Note that in this step we use
        the hybrid distance HD() instead of the original distance D()
7   for each candidate sequence S do // Perform the post-processing step.
8       Retrieve the real data sequence S from a database;
9       Identify S as a true similar sequence if D(Q,S) ≤ ε;
10  end for
11  Return the true similar data sequences;
```

(그림 3) 하이브리드 저차원 변환을 사용한 전체 매칭 알고리즘

식 (2))를 만족하는 반면에, 제안한 하이브리드 저차원 변환은 하이브리드 거리에 대해서 Parseval의 정리(즉, 공식 (4))를 만족하기 때문이다.

(그림 3)은 하이브리드 저차원 변환을 사용하는 경우의 전체 매칭 알고리즘[1]을 나타낸다. (그림 3)의 알고리즘은 참고문헌 [2, 3, 5] 등에 의해 서브시퀀스 매칭으로, 참고문헌 [6, 10] 등에 의해 k-최근접 검색(k-nearest neighbor search)으로 쉽게 확장될 수 있다. 알고리즘의 입력으로는 전체 매칭을 수행할 질의 시퀀스(Q)와 허용치(ε), 그리고 하이브리드 변환을 위한 저차원 변환들이 주어진다. 물론, 이때 주어진 저차원 변환들은 다차원 인덱스 구성에서 사용한 저차원 변환들과 동일해야 한다. (그림 3)의 알고리즘을 보면, 먼저 질의 시퀀스를 주어진 저차원 변환들을 사용하여 저차원 점으로 하이브리드 변환하는데(라인 1~4), 이 과정은 (그림 2)의 라인 2~5와 동일한 과정이다. 다음으로, 하이브리드 저차원 변환된 점과 주어진 허용치를 사용하여 범위 질의를 구성한다(라인 5). 그리고, 범위 질의를 사용하여 인덱스를 검색하여, 후보 데이터 시퀀스들을 식별해 낸다(라인 6). 이러한 인덱스 검색 과정에 있어서, 기존 전체 매칭의 경우 거리 함수로서 일반적인 L_p-거리를 사용한다. 반면에, 본 논문에서는 하이브리드 변환을 사용해야 하므로, 정의 2에서 제시한 하이브리드 거리를 사용해야 한다. 이는 다차원 인덱스에서 거리 함수를 계산하는 일부분만 변경함으로써 쉽게 구현할 수 있다. 후보 데이터 시퀀스들을 구한 이후에는 후처리 과정을 수행한다(라인 7~10). 즉, 각 후보 시퀀스에 대해서 데이터베이스(디스크)에 저장된 데이터 시퀀스를 인출한 후(라인 8), 거리 계산을 통하여 착오 해답(false alarms 혹은 false positives)인지 실제 유사 시퀀스인지의 여부를 판단한다(라인 9). 마지막으로, 실제 유사 데이터 시퀀스만을 결과로서 반환한다(라인 11).

5. 성능 평가

5.1 실험 환경 및 데이터

제안한 하이브리드 저차원 변환이 여러 시계열 데이터에 잘 적용됨을 보이기 위하여, 본 논문에서는 특징이 다른 네 가지 종류의 데이터를 실험에 사용하였다. 첫 번째 데이터는 329,112개 엔트리로 구성된 실제 주식 데이터[2, 3]로서, 이를 *STOCK-DATA*라 한다. 두 번째 데이터는 주식 데이터와 유사한 특성을 갖는 100만개 엔트리의 랜덤 워크 데이터(random walk data)[2, 3, 5]로서, 이 데이터를 *WALK-DATA*라 한다. 세 번째 데이터는 사인(sine) 함수를 사용하여 생성한 100만개의 스트리밍 시계열[15, 18]로서, 이를 *SINE-DATA*라 한다. 네 번째 데이터는 유사 주기가 반복하여 나타나는 100만개 엔트리의 합성 데이터[3, 5]로서, 이를 *PERIOD-DATA*라 한다.

저차원 변환으로는 DFT[1, 2, 5]와 PAA[8, 16]의 두 변환만을 사용하였는데, 이는 DCT는 DFT와, DWT는 PAA와 각각 특징이 유사하기 때문이다. 그리고, 제안한 하이브리드 저차원 변환은 이들 두 변환을 통합하는 방식으로 구현하였으며, 특성은 DFT 및 PAA 변환 각각에서 네 개를 추출하여 사용하였다. 유사 시퀀스 매칭 방법으로는 서브시퀀스 매칭 방법 중의 하나인 DualMatch[3]를 사용하였다. 성능평가는 하이브리드 저차원 변환과 DFT 혹은 PAA의 단일 저차원 변환을 비교하는 방식을 취하였다. 실험에서 E-Index와 비교 실험은 수행하지 않았는데, 이는 E-Index와 하이브리드 접근법은 문제 해결의 목적이 다르기 때문이다. 즉, E-Index는 최적의 성능을 목적으로, 미리 분석/예측한 데이터를 바탕으로 최적의 저차원 변환을 선택하고 여러 다차원 인덱스를 사용하는 방법을 취하였다. 반면에, 제안한 접근법은 실제 사용자가 데이터에 대한 지식이나 분석/예측 없이도 실용적으로 사용할 수 있도록 하기 위하여, 여러 저차원 변환을 하나의 인덱스에서 통합하여 사용하는 방법을 취하였다. 이와 같이, 문제 해결의 목적이 다른 두 방법의 성능을 직접적으로 비교는 큰 의미가 없다고 판단하였기 때문이다.

실험 결과로는 질의 시퀀스 길이를 256으로 고정(윈도우 크기를 128으로 고정)하고, 선택율(selectivity)[2, 3]을 1.0E-05, 1.0E-04, 1.0E-03으로 달리하면서 각 변환에 의한 서브시퀀스 매칭의 실행 시간을 측정하였다. 질의 시퀀스 길이를 256으로 고정한 이유는 실험 결과가 질의 시퀀스 길이 자체에는 큰 영향을 받지 않기 때문이다. 즉, 현재 길이인 256에서 4개 특성을 추출하였는데, 길이를 512로 변경할 경우 특성을 8개 추출하면 매우 유사한 결과를 얻을 수 있으므로, 질의 시퀀스 길이는 256으로 고정하고 실험을 수행하였다. 다음으로, 선택율 범위를 1.0E-05, 1.0E-04, 1.0E-03의 세 가지만 사용한 이유는 일반적으로 선택율이 낮은 경우(1.0E-05~1.0E-03)가 높은 경우(1.0E-03~1.0E-01)보다 자주 사용되기 때문이다[2, 5]. 또한, 선택율이 높은 경우를 실험하면 저차원 변환 종류에 따른 성능 차이가 거의 없어지는 것으로 나타나는데, 이는 선택율이 높아져, 1.0E-01에 가까워지면 대다수의 서브시퀀스가 후보로 선택되기 때문이다. 이에 따라, 본 논문에서는 선택율 범위가 낮은 1.0E-05, 1.0E-04, 1.0E-03의 세 가지를 실험에 사용하였다. 질의 시퀀스는 데이터 시퀀스의 랜덤 오프셋(random offset)을 시작 엔트리로 하는 서브시퀀스를 추출하여 사용하였으며, 노이즈를 피하

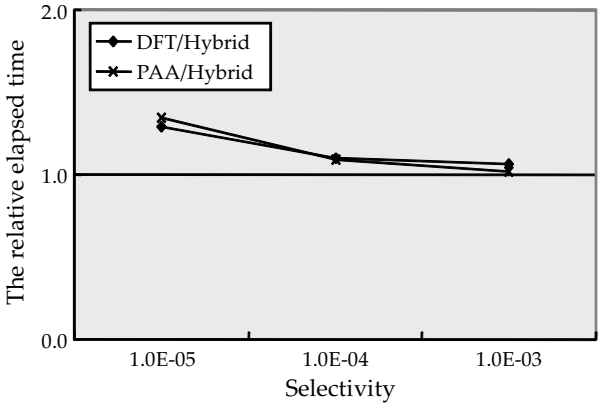
기 위하여 같은 길이를 갖는 10개의 다른 질의 시퀀스에 대해서 실험한 후 평균을 취한 값을 실험 결과로 하였다. 실험을 수행한 하드웨어 플랫폼은 Intel Pentium IV 2.80GHz, 512MB RAM, 70.0GB 하드디스크를 장착한 PC이다. 그리고, 소프트웨어 플랫폼은 GNU/Linux Version 2.6.6 운영체제이다.

5.2 실험 결과

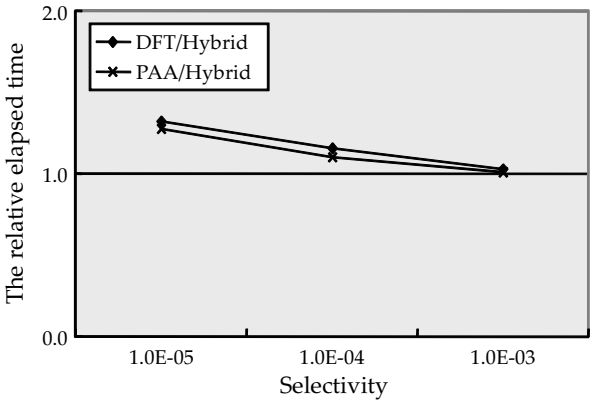
(그림 4)는 STOCK-DATA에 대한 실험 결과를 나타낸다. 그림에서 가로축은 선택율을, 세로축은 하이브리드 저차원 변환에 대한 DFT와 PAA의 상대적 실행 시간 비율을 나타낸다. (그림 4)를 보면, DFT와 PAA는 거의 유사한 성능을 보인 반면에, 제안한 하이브리드 변환은 DFT 및 PAA에 비해 우수한 성능을 보임을 알 수 있다. 이는 하이브리드 변환이 시계열 데이터의 모든 구간에 대해서 DFT와 PAA를 통합하는 최적의 변환을 수행하기 때문이다. 즉, 각 부분 부분에서는 DFT나 PAA가 최적일 수 있으나, 시계열 전체 데이터 측면에서는 DFT와 PAA를 통합한 하이브리드 변환이 가장 좋은 성능을 나타냄을 의미한다. (그림 4)에서 PAA가 DFT보다 약간 우수한 결과를 보이는데, 이는 STOCK-DATA의 경우 다른 데이터에 비해서 이웃한 엔트리들의 변화가 비교적 크고, 이는 PAA의 평균 방식이 DFT의 계수 방식 보다 유리하기 때문이다. 즉, 이웃한 엔트리들의 변화가 클 경우, PAA는 구간별 평균을 비교적 정확히 계산해내나, DFT는 에너지가 집중된 처음 몇 개의 특성을 추출하기 어렵기 때문이다. (그림 4)에서 선택율이 높아질 수록 저차원 변환에 따른 성능 차이가 줄어듦을 알 수 있다. 이는 선택율이 높아질 수록 저차원 변환을 사용하는 인덱싱 과정보다는 실제 데이터 시퀀스를 액세스하는 후처리 과정이 성능에 큰 영향을 미치기 때문이다. (그림 4)의 STOCK-DATA 실험 결과를 요약하면, 제안한 하이브리드 저차원 변환은 DFT에 비해 최대 32.4%, PAA에 비해 최대 27.1%까지 성능을 향상시킨 것으로 나타났다.

(그림 5)는 WALK-DATA에 대한 실험 결과를 나타낸다. (그림 5)의 WALK-DATA에 대한 결과는 (그림 4)의 STOCK-DATA에 대한 결과와 매우 유사함을 알 수 있다. 이는 WALK-DATA가 주식 데이터를 모델링[2]하여, STOCK-DATA와 그 성격이 유사하기 때문이다. (그림 5)의 결과를 보면, 선택율이 낮은 구간에서는 DFT가 PAA에 비해 우수함을 알 수 있는데, 이는 STOCK-DATA에 비해 이웃한 엔트리 변화가 작아 DFT에서 적은 특성에 에너지 집중이 보다 잘되기 때문이다. 반면에, 선택율이 높은 구간에서는 오히려 PAA가 DFT보다 우수한 것으로 나타났는데, 이는 선택율이 높아져 많은 유사한 시퀀스를 매칭 대상으로 할 경우, 각 구간별로 특성을 추출하는 PAA가 시퀀스 전체에서 특성을 추출하는 DFT에 비해 시퀀스 간 거리 구분 성능이 뛰어나기 때문이다. (그림 5)의 WALK-DATA의 실험 결과를 요약하면, 제안한 하이브리드 변환이 DFT에 비해 최대 28.9%, PAA에 비해 최대 34.5%까지 성능을 향상시킨 것으로 나타났다.

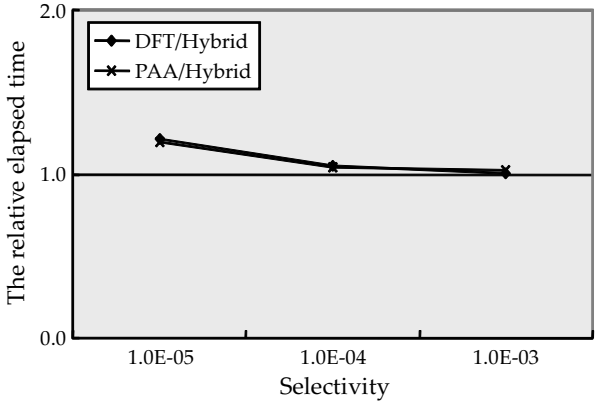
(그림 6)은 SINE-DATA에 대한 실험 결과를 나타낸다. (그림 4) 및 (그림 5)와 마찬가지로, 가로축은 선택율을, 세로축은 실행 시간의 상대적 비율을 나타낸다. 전체적인 경향은 (그림 4) 및 (그림 5)와 유사함을 알 수 있다. 즉, 하이브리드 변환이 DFT나 PAA에 비해 우수한 성능을 보이고 있다. 이는 제안한 하이브리드 변환이 DFT와 PAA의 특징



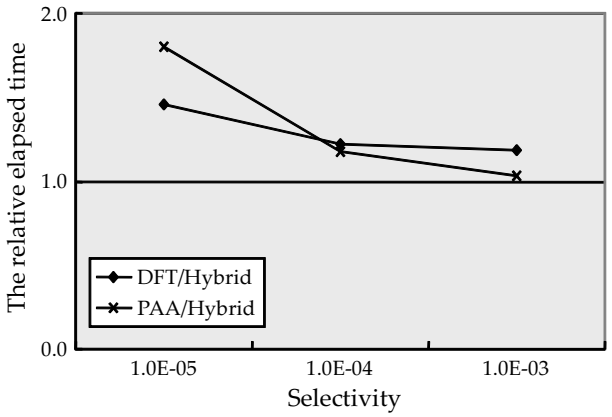
(그림 5) WALK-DATA에 대한 실험 결과



(그림 4) STOCK-DATA에 대한 실험 결과



(그림 6) SINE-DATA 에 대한 실험 결과



(그림 7) PERIOD-DATA에 대한 실험 결과

을 통합하여 발휘하기 때문이다. (그림 6)의 결과를 보면, 하이브리드 변환이 가장 우수한 반면에 DFT와 PAA는 성능 차이가 거의 없는 것으로 나타났다. 이는 SINE-DATA가 이웃한 엔트리 변화가 적은 WALK-DATA를 모태로 생성된 반면에, 주기성을 가지므로 유사한 시퀀스가 반복하여 나타나는 특징 때문이다. 즉, 엔트리 변화가 작은 점은 DFT에 유리하게 작용하나, 유사 시퀀스가 반복하여 나타나는 성질은 PAA에 유리하게 작용하여, 두 방법에 의한 차이가 거의 비슷하게 나타났다. (그림 6)의 SINE-DATA의 결과에서 하이브리드 변환은 DFT에 비해 최대 21.0%, PAA에 비해 최대 19.3%까지 성능을 향상시킨 것으로 나타났다.

(그림 7)은 PERIOD-DATA에 대한 실험 결과를 나타낸다. 지금까지의 실험 결과와 유사하게 하이브리드 저차원 변환이 DFT나 PAA에 비해 우수한 성능을 보이고 있다. (그림 7)을 보면 일부 선택율 구간에서는 DFT가 PAA보다, 다른 일부 구간에서는 PAA가 DFT보다 우수한 성능을 보이고 있다. 정확히 설명하면, 선택율이 낮은 경우에는 DFT가, 선택율이 높은 구간에서는 PAA가 우수한 성능을 보이고 있다. 이는 앞서의 WALK-DATA 및 SINE-DATA의 실험 결과와 같은 이유로 해석할 수 있다. 즉, 선택율이 낮은 구간에서는 엔트리 변화의 효과가 크게 발휘된 반면에, 선택율이 높은 구간에서는 PERIOD-DATA의 특징인 유사 시퀀스의 반복 효과가 크게 발휘되기 때문이다. 엔트리 변화의 효과가 크게 발휘된 선택율이 낮은 구간에서는 DFT가 우수한 성능을 보이고, 유사 시퀀스의 반복 효과가 크게 발휘된 선택율이 높은 구간에서는 오히려 PAA가 우수한 성능을 보인 것이다. 이와 같이 선택율에 따라 최적의 인덱싱을 보이는 저차원 변환이 달라질 수 있다. 그러나, DFT와 PAA의 특징을 통합한 하이브리드 저차원 변환은 이들 모든 구간에 있어서 가장 우수한 성능을 보이고 있다. 결과적으로, 제안한 하이브리드 변환은 데이터의 종류나 선택율의 범위에 관계없이 단일 저차원 변환보다 우수한 성능을 보인다고 말할 수 있다. (그림 7)의 PERIOD-DATA의 실험 결과에서는 하이브리드 변환이 DFT에 비해 최대 45.6%, PAA에 비해 최대 79.7%까지 성능을 향상시킨 것으로 나타났다.

지금까지의 실험 결과를 종합하면, 하이브리드 저차원 변환은 시계열 데이터의 종류나 선택율 범위에 관계없이 단일 저차원 변환을 사용하는 경우보다 우수한 성능을 보인다고 할 수 있다. 가장 좋은 결과를 보인 경우는 선택율 1.0E-05로 낮고 PERIOD-DATA를 사용한 경우로서, 이때 하이브리드 변환은 DFT 보다는 최대 45.6%까지, PAA 보다는 최대 79.7%까지 성능을 향상시키는 것으로 나타났다. 가장 좋지 않은 결과를 보인 경우는 선택율이 1.0E-03으로 높고 SINE-DATA를 사용한 경우인데, 이때 역시 하이브리드 변환은 DFT에 비해 최대 21.0%, PAA에 비해 최대 19.3%까지 성능을 나타낸 것으로 나타났다. 이 같은 결과를 종합하면, 데이터나 선택율에 따라 어떤 저차원 변환을 선택할 고민이 없이, 여러 저차원 변환을 통합하는 하이브리드 변환을 사용하면 유사 시퀀스 매칭에 있어서 보다 향상된 성능을 얻을 수 있음 의미한다. 또한, 비록 하이브리드 변환이 기존 저차원 변환에 비해 뛰어난(몇 배 혹은 몇 십배) 성능 개선 효과를 거두지는 못하지만, 이들 여러 변환의 장점을 취할 수 있는 우수한 접근법이라 할 수 있다. 즉, 새로운 저차원 변환이 개발되어 우수한 성능을 보인다면, 이를 본 논문의 하이브리드 접근법에 적용하면 보다 더 우수한 저차원 변환 개발이 가능한 프레임워크를 제공한다고 말할 수 있다.

6. 결 론

본 논문에서는 하이브리드 저차원 변환의 개념을 제시하고, 이를 사용한 유사 시퀀스 매칭 방법을 제안하였다. 제안한 하이브리드 저차원 변환은 하나의 시계열 데이터에 여러 저차원 변환을 동시에 적용하는 방법으로서, 하나의 인덱스에서 두 개 이상의 저차원 변환들을 통합하여 사용하도록 하였다. 이를 통하여 여러 저차원 변환의 특성을 동시에 발휘할 수 있고, 궁극적으로 유사 시퀀스 매칭의 성능을 향상시킬 수 있다.

본 논문의 공헌은 다음과 같이 요약할 수 있다. 첫째, 여러 개의 저차원 변환을 사용하여 하나의 시퀀스에서 서로 다른 특징을 지닌 여러 특성들을 추출하는 하이브리드 저차원 변환 개념을 제안하였다. 둘째, 하이브리드 변환된 두 시퀀스간의 거리를 계산하기 위한 하이브리드 거리를 제시하고, 이를 사용하면 유사 시퀀스 매칭을 정확하게 수행할 수 있음을 정리 1에서 정형적으로 증명하였다. 셋째, 제안한 하이브리드 저차원 변환과 하이브리드 거리를 사용하는 인덱스 구성 및 유사 시퀀스 매칭 알고리즘을 제시하였다. 넷째, 제안한 하이브리드 접근법이 단일 저차원 변환을 사용하는 경우에 비해서 우수한 성능을 보임을 실험을 통하여 확인하였다. 이 같은 결과를 볼 때, 제안한 하이브리드 접근법은 다양한 특성을 지닌 다양한 시계열 데이터에 두루 적용될 수 있는 우수한 방법이라 사료된다.

참 고 문 헌

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

[2] Faloutsos, C., Ranganathan, M., and Manolopoulos, Y., "Fast Subsequence Matching in Time-Series Databases," *In Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, Minneapolis, Minnesota, pp.419-429, May, 1994.

[3] Moon, Y.-S., Whang, K.-Y., and Loh, W.-K., "Duality-Based Subsequence Matching in Time-Series Databases," *In Proc. the 17th Int'l Conf. on Data Engineering*(ICDE), IEEE, Heidelberg, Germany, pp.263-272, April, 2001.

[4] Keogh, E. J. et al., "LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures," *In Proc. Int'l Conf. on Very Large Data Bases (VLDB)*, Seoul, Korea, pp.882-893, Sept., 2006.

[5] Moon, Y.-S., Whang, K.-Y., and Han, W.-S., "General Match: A Subsequence Matching Method in Time-Series Databases Based on Generalized Windows," *In Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, Madison, Wisconsin, pp.382-393, June, 2002.

[6] Chan, K.-P., Fu, A. W.-C., and Yu, C. T., "Haar Wavelets for Efficient Similarity Search of Time-Series: With and Without Time Warping," *IEEE Trans. on Knowledge and Data Engineering*, Vol.15, No.3, pp.686-705, Jan./Feb., 2003.

[7] Loh, W.-K., Kim, S.-W., and Whang, K.-Y., "A Subsequence Matching Algorithm that Supports Normalization Transform in Time-Series Databases," *Data Mining and Knowledge Discovery*, Vol.9, No.1, pp.5-28, July, 2004.

[8] Keogh, J., Chakrabarti, K., Mehrotra, S., and Pazzani, M. J., "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases," *In Proc. of Int'l Conf. on Management of Data*, ACM SIGMOD, Santa Barbara, CA, pp.151-162, May, 2001.

[9] Keogh, J., Chakrabarti, K., Pazzani, M. J., and Mehrotra, S., "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," *Knowledge and Information Systems*, Vol.3, No.3, pp.263-286, Aug., 2001.

[10] Keogh, E. J., Chu, S., and Pazzani, M. J., "Ensemble-Index: A New Approach to Indexing Large Databases," *In Proc. of the 7th Int'l Conf. on Knowledge Discovery and Data Mining*, ACM SIGKDD, San Francisco, CA, pp.117-125, Aug., 2001.

[11] Berchtold, S., Bohm, C., and Kriegel, H.-P., "The Pyramid-Technique: Towards Breaking the Curse of Dimensionality," *In Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, Seattle, Washington, pp.142-153, June, 1998.

[12] Beckmann, N., Kriegel, H.-P., Schneider, R., and Seeger, B., "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles," *In Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, Atlantic City, New Jersey, pp.322-331, May, 1990.

[13] Lim, S.-H., Park, H.-J., and Kim, S.-W., "Using Multiple Indexes for Efficient Subsequence Matching in Time-Series Databases," *In Proc. of the 11th Int'l Conf. on Database Systems for Advanced Applications (DASFAA 2006)*, Singapore, pp.65-79, Apr., 2006.

[14] Moon, Y.-S. and Kim, J., "A Single Index Approach for Time-Series Subsequence Matching that Supports Moving Average Transform of Arbitrary Order," *In Proc. of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2006)*, Singapore, pp.739-749, Apr., 2006.

[15] Moon, Y.-S., "An MBR-Safe Transform for High-Dimensional MBRs in Similar Sequence Matching," *In Proc. of the 12th Int'l Conf. on Database Systems for Advanced Applications (DASFAA 2007)*, Bangkok, Thailand, pp.79-90, April, 2007.

[16] Yi, B.-K. and Faloutsos, C., "Fast Time Sequence Indexing for Arbitrary Lp Norms," *In Proc. of the 26th Int'l Conf. on Very Large Data Bases*, Cairo, Egypt, pp.385-394, Sept., 2000.

[17] Hsieh, M. J., Chen, M. S., and Yu, P. S., "Integrating DCT and DWT for Approximating Cube Streams," *In Proc. of the 14th ACM Int'l Conf. on Information and Knowledge Management*, Bremen, Germany, pp.179-186, Oct., 2005.

[18] Gao, L. and Wang, X. S., "Continually Evaluating Similarity-based Pattern Queries on a Streaming Time Series," *In Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, Madison, Wisconsin, pp.370-381, June, 2002.

문 양 세

e-mail : ysmoon@kangwon.ac.kr

1991년 2월 한국과학기술원 과학기술대학 전산학과(학사)

1993년 2월 한국과학기술원 전산학과(석사)

2001년 8월 한국과학기술원 전자전산학과 전산학전공 (박사)

1993년 2월～1997년 2월 현대전자산업(주) 통신사업본부 주임연구원

2001년 9월～2002년 2월 ㈜현대시스콤 호처리개발실 선임연구원

2002년 2월～2005년 2월 ㈜인프라밸리 기술연구소 기술위원(이사)

2005년 3월～현 재 한국과학기술원 첨단정보기술연구센터 연구원

2005년 3월～현 재 강원대학교 IT특성화대학 컴퓨터과학전공 조교수

관심분야 : Data Mining, Knowledge Discovery, Stream Data, Storage System, Database Applications, Mobile/Wireless Communication Services & Systems

김 진 호

e-mail : jhkim@kangwon.ac.kr

1982년 2월 경북대학교 전자공학과(학사)

1985년 2월 한국과학기술원 전산학과(석사)

1990년 2월 한국과학기술원 전산학과(박사)

1995년 8월~1996년 7월 미국 미시간 대학교 객원 교수

2003년 2월~2004년 2월 미국 Drexel University 객원 교수

1999년 3월~현 재 한국과학기술원 첨단정보기술연구센터 연구원

1990년 8월~현 재 강원대학교 IT특성화대학 컴퓨터과학전공 교수

2006년 8월~현 재 강원대학교 중앙교육연구전산원 원장

관심분야 : Data warehouse, OLAP, Data Mining, Real-time/Embedded Database, Main-memory database, Data Modeling, Web Database Technology