MBR-Safe 변환: 유사 시퀀스 매칭에서 고차원 MBR의 저차원 변환

(MBR-Safe Transform: Lower-Dimensional Transformation of High-Dimensional MBRs in Similar Sequence Matching)

문 양 세 †

(Yang-Sae Moon)

요 약 대부분의 유사 시퀀스 매칭 방법은 다차원 색인을 사용한 검색 속도의 향상을 위해, 많은 수의 고차원 시퀀스를 저차원 변환한 후 이들 변환된 시퀀스들을 포함하는 저차원 MBR을 구성한다. 본 논문에서는 고차원 MBR 자체를 직접 저차원 MBR로 변환하는 정형적인 방법을 제안하고, 이를 사용하면 유사 시퀀스 매칭에서 필요한 저차원 변환 횟수를 획기적으로 줄일 수 있음을 보인다. 이를 위해, 우선 변환의 *MBR-safe* 개념을 정형적으로 제안한다. 어떤 변환이 MBR-safe하다 함은 고차원 MBR을 직접 변환한 저차원 MBR이 개별 고차원 시퀀스가 변환된 저차원 시퀀스를 모두 포함함을 의미한다. 다음으로, 기존 저차원 변환 중에서 가장 널리 사용되는 DFT와 DCT에 대해 각각 MBR-safe 변환을 제안한다. 먼저, 기존 DFT와 DCT가 MBR-safe하지 않음을 보이고, DFT와 DCT를 확장한 *mbrDFT*와 *mbrDCT*를 각각 정의한다. 그리고, 이들 mbrDFT와 mbrDCT가 MBR-safe함을 정형적으로 증명한다. 또한, mbrDFT(혹은 mbrDCT)가 고차원 MBR을 저차원 MBR로 직접 변환하는 DFT(혹은 DCT) 기반의 최적 MBR-safe 변환임을 증명한다. 분석과 실험 결과, 제안한 mbrDFT 및 mbrDCT를 사용하면 저차원 변환 횟수를 획기적으로 줄이고 성능을 크게 향상 시킨 것으로 나타났다. 이 같은 결과를 볼 때, 본 논문에서 제시한 MBR-safe 개념은 고차원 MBR의 저차원 변환이 필요한 많은 응용에 활용될 수 있는 유용한 연구 결과라 사료된다.

키워드 : MBR-safe 변환, 저차원 변환, MBR-safe, 데이타 마이닝, 유사 시퀀스 매칭

Abstract To improve performance using a multidimensional index in similar sequence matching, we transform a high-dimensional sequence to a low-dimensional sequence, and then construct a low-dimensional MBR that contains multiple transformed sequences. In this paper we propose a formal method that transforms a high-dimensional MBR itself to a low-dimensional MBR, and show that this method significantly reduces the number of lower-dimensional transformations. To achieve this goal, we first formally define the new notion of *MBR-safe*. We say that a transform is MBR-safe if a low-dimensional MBR to which a high-dimensional MBR is transformed by the transform contains every individual low-dimensional sequence to which a high-dimensional sequence is transformed. We then propose two MBR-safe transforms based on DFT and DCT, the most representative lower-dimensional transformations. For this, we prove the traditional DFT and DCT are not MBR-safe, and define new transforms, called *mbrDFT* and *mbrDCT*, by extending DFT and DCT, respectively. We also formally prove these mbrDFT and mbrDCT are MBR-safe. Moreover, we show that mbrDFT(or mbrDCT) is optimal among the DFT-based(or DCT-based) MBR-safe transforms that directly convert a high-dimensional MBR itself into a low-dimensional MBR. Analytical and experimental results show that the proposed mbrDFT and mbrDCT reduce the number of lower-dimensional transformations drastically, and improve performance significantly compared with the naïve transforms. These results indicate that our MBR-safe transforms provides a useful framework for a variety of applications that require the lower-dimensional transformation of high-dimensional MBRs.

Key words : MBR-safe Transform, Lower-Dimensional Transformations, Data Mining, Similar Sequence Matching

---

· 본 연구는 첨단정보기술연구센터를 통하여 과학기술부/한국과학재단의 지원을 받았음

† 정 회 원 : 강원대학교 컴퓨터학부 컴퓨터과학전공 교수

ysmoon@kangwon.ac.kr

논문접수 : 2006년 7월 26일

심사완료 : 2006년 9월 6일

1. 서 론

시계열 데이타(time-series data)란 각 시간별로 측정한 실수 값의 시퀀스로, 그 예로는 주식 데이타, 환율 데이타, 날씨 변동 데이타 등이 있다[1-4]. 시계열 데이타베이스에 저장된 시계열 데이타를 데이타 시퀀스라 부르며, 사용자에 의해 주어진 시퀀스를 질의 시퀀스라 부른다. 그리고, 주어진 질의 시퀀스와 유사한 데이타 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 함수($=L_2$)를 비롯하여, 맨하탄 거리($=L_1$), 최대 거리($=L_\infty$) 등의 L_p-거리 함수($=\sqrt[p]{\sum_{i=0}^{n-1}|x_i-y_i|^p}$)를 주로 사용한다[1,2,6-8].

대부분의 유사 시퀀스 매칭 방법은 고차원인 시퀀스를 다차원 색인에 저장하기 위하여 **저차원 변환**(*lower-dimensional transformation*)을 사용한다[1,2,5-9]. 이와 같이 저차원 변환을 수행하는 이유는 다차원 색인의 고차원 문제(high dimensionality problem)[10]로 인하여, 고차원인 시퀀스를 R*-트리[11]와 같은 다차원 색인에 직접 저장하기 어렵기 때문이다. 즉, 유사 시퀀스 매칭에서 검색 속도의 향상을 위해서 다차원 색인을 사용하는데, 이때 고차원 문제를 피하고, 색인의 저장 공간을 줄이기 위하여 저차원 변환을 사용하는 것이다. 이러한 저차원 변환은 시계열 데이타베이스에서 유사 시퀀스 매칭 문제를 처음 소개한 Agrawal 등[1]의 전체 매칭(whole matching) 이래, 여러 전체 매칭[6,12,13]은 물론 서브시퀀스 매칭(subsequence matching)[2,7-9,14]에서 폭넓게 사용되었다. 또한, 최근 스트리밍 시계열에서의 유사 시퀀스 매칭[15,16]에서도 질의 시퀀스 혹은 스트리밍 시계열의 차원을 줄이기 위하여 사용되고 있다.

본 논문에서는 이러한 기존 유사 시퀀스 매칭에서 MBR(minimum bounding rectangle)을 구성하는 방법에 주목한다. 기존 유사 시퀀스 매칭 방법에서는 다차원 색인에 저장되는 점의 개수를 줄이기 위하여 MBR을 사용한다. 즉, 저차원 변환된 여러 점을 포함하는 MBR을 구성하고, 이들 개별 점 대신 MBR만을 색인에 저장하는 것이다. 예를 들어, 서브시퀀스 매칭에서는 데이타 시퀀스를 나눈 윈도우들을 저차원 변환한 여러 점들을 포함하는 MBR을 구성하거나[2,8], 질의 시퀀스를 나눈 윈도우들을 저차원 변환한 여러 점들을 포함하는 MBR을 구성한다[5,7,14]. 결국, 기존의 MBR 구성 방법은 수십~수천 개[2,14]의 고차원 시퀀스(혹은 윈도우) 각각을 저차원 시퀀스(점)로 변환한 후, 이들 저차원 시퀀스들을 대상으로 MBR을 구성한다. 이에 따라, 수많은 저차원 변환이 수행되는데, 본 논문에서는 이러한 저차원 변환의 횟수를 줄이는 방법을 연구한다.

저차원 변환의 횟수를 줄이기 위하여, 본 논문에서는 고차원 MBR을 저차원 변환하는 방법을 제안한다. 즉, 여러 개의 고차원 시퀀스를 포함하는 고차원 MBR을 구성하고, 이 MBR 자체를 저차원 변환하는 방법이다. 이를 위해 우선 변환의 *MBR-safe* 개념을 제안한다. 어떤 변환 T가 MBR-safe하다 함은, 여러 시퀀스를 포함하는 MBR M이 T에 의해 MBR M^T로 변환되었을 때, M에 포함된 개별 시퀀스 X를 T로 변환한 X^T 역시 M^T에 모두 포함됨을 의미한다. 즉, 개별 시퀀스 X가 변환된 각 시퀀스 X^T가 MBR M이 변환된 M^T에 모두 포함되면, 변환 T는 MBR-safe하다고 정의한다. 이러한 MBR-safe 개념을 사용하면, 개별 시퀀스들을 변환하여 저차원 MBR을 구성하는 대신, 고차원 MBR 자체를 변환하여 저차원 MBR을 구성할 수 있다. 그리고, 이를 통해 저차원 MBR 구성에 필요한 저차원 변환 횟수를 크게 줄일 수 있다.

본 논문에서는 기존의 저차원 변환 중에서 가장 많이 사용되는 DFT(Discrete Fourier Transform)와 DCT(Discrete Cosine Transform)에 대해서 각각 MBR-safe 변환을 제안한다. 우선, 기존 DFT가 MBR-safe하지 않음을 보인다. 그런 다음, MBR을 고려하여 기존 DFT를 *mbrDFT*로 확장하여 정의하고, 이 mbrDFT가 MBR-safe함을 정형적으로 증명한다. 또한, 기존 DCT가 MBR-safe하지 않음을 보인다. 마찬가지로, 기존 DCT를 *mbrDCT*로 확장하여 정의하고, 이 mbrDCT가 MBR-safe함을 정형적으로 증명한다. 다음으로, 제안한 mbrDFT(혹은 mbrDCT)가 고차원 MBR을 저차원 MBR로 직접 변환하는 DFT(혹은 DCT) 기반의 최적 MBR-safe 변환임을 증명한다.

본 논문에서는 분석과 실험을 통해 MBR-safe 변환의 우수성을 보인다. 우선, 기존 저차원 변환과 MBR-safe 변환의 계산 복잡도를 유도함으로써, 제안한 MBR-safe 변환의 우수성을 분석적으로 설명한다. 다음으로, 실험을 통해 제안한 mbrDFT 및 mbrDCT를 사용하면 저차원 변환 횟수를 크게 줄일 수 있음을 보인다. 또한, 이를 통하여 제안한 MBR-safe 변환이 기존 저차원 변환에 비해 성능을 크게 향상 시킬 수 있음을 보인다. 그리고, 제안한 MBR-safe 변환에 의해 구성된 저차원 MBR과 기존 변환에 의해 구성된 저차원 MBR의 차이를 분석한다.

본 논문의 구성은 다음과 같다. 제2장에서는 유사 시퀀스 매칭 및 저차원 변환의 관련 연구를 설명한다. 제3장에서는 본 논문에서 해결하고자 하는 문제, 즉 변환의 MBR-safe 개념을 정의한다. 제4장에서는 DFT와 DCT

에 대해 두 가지 MBR-safe 변환을 제안한다. 제5장에서는 제안한 MBR-safe 변환의 계산 복잡도를 분석한다. 제6장에서는 실험을 통해 제안한 방법의 우수성을 보인다. 마지막으로, 제7장에서 결론을 맺는다.

2. 관련 연구

유사 시퀀스 매칭은 크게 전체 매칭과 서브시퀀스 매칭의 두 가지로 구분한다[2]. 전체 매칭은 질의 시퀀스와 유사한 데이타 시퀀스를 찾는 문제로서, 질의 시퀀스와 데이타 시퀀스의 길이가 동일한 특징을 갖는다[1,6,12,13]. 반면에, 서브시퀀스 매칭은 데이타 시퀀스에 포함된 서브시퀀스들 중에서 질의 시퀀스와 유사한 서브시퀀스를 찾는 문제로서, 사용자는 임의 길이의 시퀀스를 질의 시퀀스로 사용할 수 있다[2,5,7,8,14]. 또한, 유클리디안 거리 함수가 갖는 문제점을 보완하기 위하여, 이동평균[9,17], 쉬프팅 및 스케일링[18,19], 정규화[8,17], 타임 워핑[13,20,21] 등의 다양한 변환 기법이 사용되었다. 그리고, 이러한 유사 시퀀스 매칭 방법 대부분에서는 다차원 색인 사용을 위하여 저차원 변환을 사용한다.

기존의 유사 시퀀스 매칭 방법에서는 다차원 색인에 저장되는 점의 개수를 줄이기 위하여, 저차원 변환된 여러 개의 점을 포함하는 MBR을 구성하고, 이들 여러 점들을 대표하는 MBR만을 색인에 저장하거나 검색에 사용한다. 예를 들어, 참고문헌 [2,8,9]에서는 데이타 시퀀스를 나눈 윈도우들을 저차원 변환한 여러 점들을 포함하는 MBR을 구성하여 다차원 색인에 저장하는 방법을 사용하였다. 또한, 참고문헌 [5,7,14]에서는 질의 시퀀스를 나눈 윈도우들을 저차원 변환한 여러 점들을 포함하는 질의 MBR을 구성하여 다차원 색인을 검색하는 방법을 사용하였다. 그리고, 스트리밍 시계열에서 연속질의에 대한 유사 시퀀스 매칭이서도 여러 시퀀스들을 대상으로 저차원 MBR을 구성하는 방법을 사용하였다[16]. 그러나, 이들 모든 방법은 개별 고차원 시퀀스를 저차원 변환하고, 변환된 여러 저차원 시퀀스(점)들을 대상으로 MBR을 구성하였으며, 따라서 본 논문에서 제시하는 고차원 MBR 자체를 변환하는 방법과는 저차원 MBR 구성 방법에 있어서 큰 차이가 있다.

고차원 시퀀스의 저차원 변환 방법으로는 DFT, DCT, Wavelet 등 여러 가지 변환이 사용되었다. 우선, DFT는 참고문헌 [1,2,5,7-9] 등 많은 연구에서 사용되었다. 다음으로, DCT의 경우 멀티미디어 데이타[22,23] 혹은 데이타 스트림[24] 등에서 사용되었으며, (Haar) Wavelet 변환 역시 참고문헌 [6,14,25] 등 많은 연구에서 저차원 변환 방법으로 사용되었다. 이외에도, PAA (Piecewise Aggregate Approximation)[12], SVD(Singluar Value Decomposition)[26] 등 여러 가지 저차원 변환 방법이 제시되었다. 그러나, 이들 저차원 변환 방법은 대부분 고차원 시퀀스 혹은 고차원 이미지를 저차원 변환하는 것으로서, 고차원 MBR을 변환하는 방법에는 그대로 적용할 수 없다.

3. 문제 정의

MBR-safe 문제를 설명하기 위하여 본 논문에서 사용하는 주요 표기와 이에 대한 정의 및 의미는 표 1과 같다. 표 1의 정의에 따르면, 본 논문의 연구 목표는 차원(길이)이 n인 시퀀스 X와 MBR $[L,U]$가 주어졌을 때, $X \in [L,U]$이면 $X^T \in [L,U]^T$를 만족하는 변환 T를 찾는 것이다. 이러한 성질을 만족하는 변환을 본 논문에서는 MBR-safe하다 하고 다음과 같이 정형적으로 정의한다.

정의 1. 차원이 n인 시퀀스 X와 같은 차원인 MBR $[L,U]$가 주어졌을 때, 어떤 변환 T가 있어 $X \in [L,U]$이면 $X^T \in [L,U]^T$를 만족하며, T는 *MBR-safe*하다고 정의한다. 즉, 변환 T가 다음 공식 (1)을 만족하면 T는 *MBR-safe*하다고 정의한다.

$$X \in [L,U] \Rightarrow X^T \in [L,U]^T \quad (1)$$

그림 1은 MBR-safe 개념을 그림으로 나타낸 것이다. 그림에서 변환 $T1$은 MBR-safe하고, $T2$는 MBR-safe하지 않다. 변환 $T1$이 MBR-safe한 이유는 임의의 시퀀스 X가 MBR $[L,U]$에 포함된다면, $T1$에 의해 변환된 시퀀스 X^{T1} 역시 변환된 MBR $[L,U]^{T1}$(=$[\Lambda,\Upsilon]$)에 포함되기 때문이다. 반면에, 변환 $T2$의 경우는 X가 MBR $[L,U]$에 포함됨에도 불구하고 X^{T2}가 $[L,U]^{T2}$(=[A,B])에는 포함되지 않기 때문에 MBR-safe하지 않다.

표 1 주요 표기법

기호	정의/의미
X	고차원 시퀀스 혹은 고차원 윈도우 (=$\{x_0, x_1, \cdots, x_{n-1}\}$)
X^T	변환 T에 의해 변환된 (저차원) 시퀀스 (=$\{x_0^T, x_1^T, \ldots, x_{m-1}^T\}$)
$[L,U]$	고차원 MBR로서, L은 좌하점, U는 우상점을 나타냄 (=[$\{l_0, l_1, \cdots, l_{n-1}\}$, $\{u_0, u_1, \cdots, u_{n-1}\}$])
$[L,U]^T = [\wedge,\gamma]$	MBR $[L,U]$가 T에 의해 변환된 (저차원) MBR (=[$\{\lambda_0, \lambda_1, \cdots, \lambda_{m-1}\}$,$\{\upsilon_0, \upsilon_1, \cdots, \upsilon_{m-1}\}$])
$X \in [L,U]$	시퀀스 X가 MBR $[L,U]$에 포함됨을 의미함

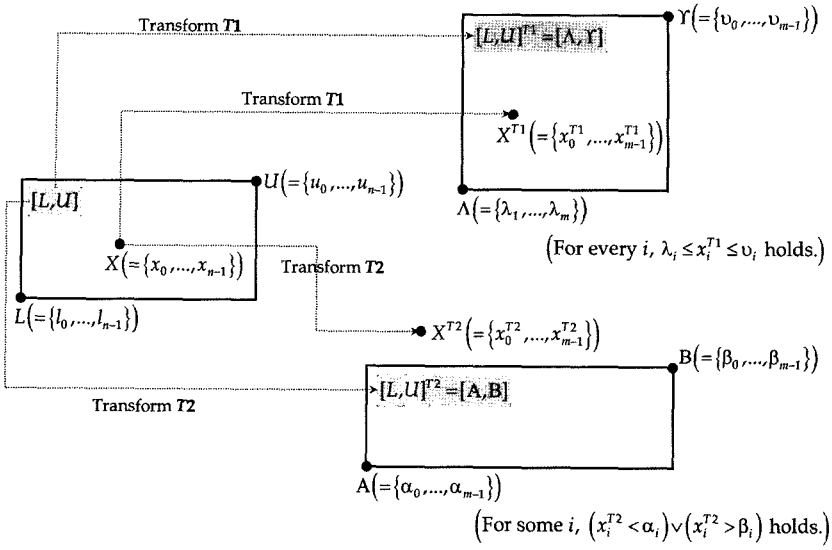
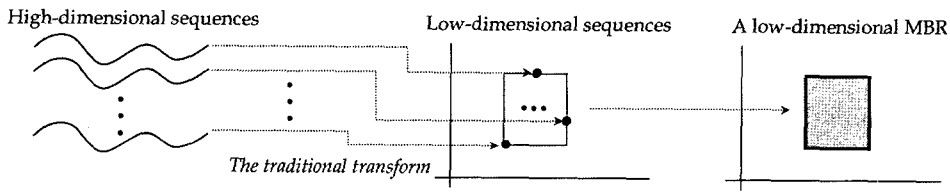
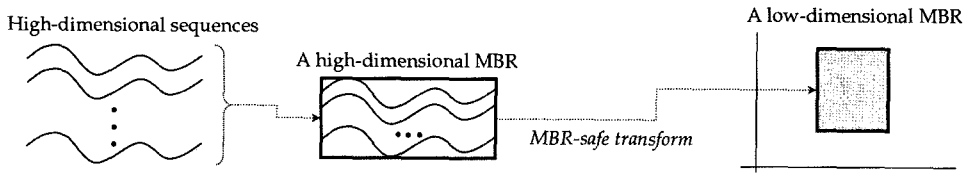


그림 1 MBR-safe한 변환($T1$)과 그렇지 않은 변환($T2$)



(a) 기존 저차원 변환에 의한 저차원 MBR 구성 방법



(b) 여러 고차원 시퀀스들에 대한 저차원 MBR 구성 방법

그림 2 여러 고차원 시퀀스들에 대한 저차원 MBR 구성 방법

MBR-safe 개념은 유사 시퀀스 매칭에서 저차원 변환 횟수를 줄이는데 사용할 수 있다. 즉, 기존 유사 시퀀스 매칭에서는 수십~수천 개의 시퀀스들 각각을 저차원 변환한 후 저차원 MBR을 구성한다[2,7,8,14]. 반면에, MBR-safe 개념을 사용하면, 수십~수천 개의 시퀀스를 포함하는 고차원 MBR을 구성한 후, 이 고차원 MBR 자체를 변환하여 저차원 MBR을 구성할 수 있다. 그림 2는 이들 두 가지의 저차원 MBR 구성 방법을 나타낸다. 그림 2(a)는 MBR-safe 변환이 아닌 기존 저차원 변환을 사용하는 예이고, 그림 2(b)는 MBR-safe 변환을 사용하는 예이다. 그림 2(a)와 같이 기존 저차원 변환을 사용하면, 수십~수천 개의 개별 시퀀스를 각각 저차원 변환한 후, 이들 변환된 점들을 포함하는 저차원 MBR을 구성한다. 반면에, 그림 2(b)와 같이 MBR-safe 변환을 사용하면, 수십~수천 개의 개별 시퀀스 대신 하나의 MBR 만을 변환하면 된다. 따라서, MBR-safe 변환을 이용하면 유사 시퀀스 매칭에서 빈번하게 발생하는 저차원 변환 횟수를 크게 줄일 수 있다.

4. MBR-Safe 변환

본 장에서는 유사 시퀀스 매칭의 저차원 변환으로 많이 사용되는 DFT와 DCT에 대해 각각 MBR-safe 변환을 제안한다. 제4.1절에서는 기존 DFT를 MBR에 적용하는 단순한 변환이 MBR-safe하지 않음을 보이고,

DFT를 확장하여 MBR-safe한 mbrDFT를 제안한다. 다음으로, 제4.2절에서는 같은 방법으로 DCT를 확장하여 mbrDCT를 제안한다.

4.1 mbrDFT: MBR-Safe DFT

DFT는 유사 시퀀스 매칭의 저차원 변환 방법으로 가장 널리 사용되었다[1,2,5,7,8,14]. 차원이 n인 시퀀스 X를 DFT로 변환한 시퀀스 Y는 다음 공식 (2)에 의해 구해진 복소수(complex number) 시퀀스 $\{y_0, y_1, ..., y_{n-1}\}$로 정의된다[1,27].

$$y_i = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t e^{-j \cdot 2\pi i t / n}, \text{ where } 0 \le i \le n-1 \quad (2)$$

그리고, 오일러의 정의[27] 및 복소수 정의에 의해 공식 (2)는 다음 공식 (3)과 같이 실수부와 허수부를 구분하여 표현할 수 있다.

$$y_i = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \cos(-2\pi i t / n) + \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \sin(-2\pi i t / n) \cdot j, \quad \text{where } 0 \le i \le n-1 \quad (3)$$

유사 시퀀스 매칭에서는 실수 시퀀스 X를 변환한 복소수 시퀀스 Y에서 에너지가 집중된 상위 몇 개의 계수(coefficient)를 선택하는 저차원 변환 방법을 사용한다. 다음 정의 2는 이와 같이 실수 시퀀스를 DFT로 변환한 후, 몇 개의 계수를 선택하는 DFT 기반의 저차원 변환을 나타낸다.

정의 2. 차원이 n인 시퀀스 X를 차원 $m(\square\, n)$으로 **DFT 저차원 변환한 시퀀스** X_{DFT}는 다음 공식 (4)에 의해 구해진 실수 시퀀스 $\{x_0^{DFT}, x_1^{DFT}, ..., x_{m-1}^{DFT}\}$로 정의한다. 그리고, 차원이 n인 MBR $[L,U]$를 차원 m으로 **DFT 저차원 변환한 MBR** $[L,U]^{DFT}$는 시퀀스 L과 U를 각각 L^{DFT}과 U^{DFT}로 변환하여 얻은 MBR $[L^{DFT}, U^{DFT}]$로 정의한다.

$$x_i^{DFT} = \begin{cases} \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \cos\theta, & i \text{ is even;} \\ \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \sin\theta, & i \text{ is odd;} \end{cases},$$

$$\text{where } \theta = -2\pi \lfloor i/2 \rfloor t / n \text{ and } 0 \le i \le m-1. \quad (4)$$

기존의 유사 시퀀스 매칭에서는 이러한 DFT 저차원 변환을 사용하여, 차원이 수십~수백인 시퀀스를 1~6 차원의 시퀀스(점)로 변환하였다.

그러나, 정의 2의 DFT 저차원 변환은 MBR-safe하지 못하다. 다음 예제 1은 DFT 저차원 변환이 MBR-safe하지 못하는 반례를 보여준다.

예제 1. 차원이 4인 시퀀스 X가 {3.00, 2.50, 3.50, 3.00}로 주어지고, 같은 차원의 MBR $[L,U]$가 L = {2.00, 1.00, 3.00, 2.00}, U = {4.00, 3.00, 5.00, 4.00}으로 주어졌다고 하자. 그러면, 주어진 X와 $[L,U]$ 사이에는 $X \in [L,U]$의 관계가 성립한다. 그리고, DFT 저차원 변환을 통해 4차원을 2차원으로 변환하려 한다. 그러면, DFT 저차원 변환의 정의에 의해, 시퀀스 X가 변환된 X^{DFT}는 {6.00, -0.25}가 되고, MBR $[L,U]$가 변환된 $[L,U]^{DFT}$는 L^{DFT} = {4.00, -0.50}, U^{DFT} = {8.00, -0.50}의 MBR이 된다.[1] 여기에서, $4.00 \le 6.00 \le 8.00$이므로 $l_0^{DFT} \le x_0^{DFT} \le u_0^{DFT}$가 성립한다. 그러나, $-0.50 \le -0.25 \not\le -0.50$이므로 $l_2^{DFT} \le x_2^{DFT} \le u_2^{DFT}$는 성립하지 않는다. 따라서, X와 $[L,U]$가 DFT 저차원 변환된 X^{DFT}와 $[L,U]^{DFT}$ 사이에는 $X^{DFT} \in [L,U]^{DFT}$의 관계가 성립하지 않는다. 다시 말해서, DFT 저차원 변환은 MBR-safe하지 않다.□

이와 같이 DFT 저차원 변환은 MBR-safe하지 않으므로, 고차원 MBR의 저차원 변환에는 사용할 수 없다. 따라서, 본 논문에서는 기존 DFT 정의를 확장하여 다음과 같이 mbrDFT를 정의한다.

정의 3. 차원이 n인 MBR $[L,U]$를 차원 $m(\ll n)$으로 **mbrDFT 변환한 MBR** $[L,U]^{mbrDFT}$는 다음 공식 (5)에 의해 구해진 MBR $[\Lambda, \Upsilon]$으로 정의한다. 그리고, 시퀀스 X를 **mbrDFT 변환한 시퀀스** X^{mbrDFT}는 X를 DFT 저차원 변환한 시퀀스 X^{DFT}와 동일하다.

$$\lambda_i = \begin{cases} \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} a_t \cos\theta, & i \text{ is even;} \\ \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} b_t \sin\theta, & i \text{ is odd;} \end{cases}, \quad \upsilon_i = \begin{cases} \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} c_t \cos\theta, & i \text{ is even;} \\ \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} d_t \sin\theta, & i \text{ is odd;} \end{cases},$$

$$\text{where } \begin{cases} a_t = l_t, c_t = u_t, & \cos\theta \ge 0; \\ a_t = u_t, c_t = l_t, & \cos\theta < 0; \\ b_t = l_t, d_t = u_t, & \sin\theta \ge 0; \\ b_t = u_t, d_t = l_t, & \sin\theta < 0; \end{cases}, \quad \theta = -2\pi \lfloor i/2 \rfloor t / n, \text{ and } 0 \le i \le m-1. \quad (5)$$

정의 3의 mbrDFT는 DFT의 MBR-safe를 보장하기 위하여, 변환된 MBR $[\Lambda, \Upsilon]$의 좌하점 Λ와 우상점 Υ이 원 MBR $[L,U]$내의 개별 시퀀스가 저차원 변환된 모든 점들을 포함하도록 유도된 것이다. 다음 정리 1은 mbrDFT가 MBR-safe함을 나타낸다.

정리 1. 시퀀스 X와 MBR $[L,U]$가 주어졌을 때, $X \in [L,U]$이면, X와 $[L,U]$를 각각 mbrDFT 변환한 X^{mbrDFT}와 $[L,U]^{mbrDFT}(=[\Lambda, \Upsilon])$ 사이에는 $X^{mbrDFT} \in [L,U]^{mbrDFT}$의 관계가 성립한다. 즉, mbrDFT는 MBR-safe하다.

증명. 모든 i에 대해서 $\lambda_i \le x_i^{mbrDFT} \le \upsilon_i$가 성립하면,

1) 실수 엔트리를 갖는 시퀀스에 대한 DFT에서는 첫 번째 복소수의 허수부(즉, x_1^{DFT})가 항상 0이 된다. 따라서, 본 논문에서는 이 값(x_1^{DFT})을 제외하고, x_0^{DFT}와 x_2^{DFT}의 두 개 차원을 사용하였다.

$X^{mbrDFT} \in [\Lambda,\Upsilon]=[L,U]^{mbrDFT}$ 가 성립한다. 우선, i가 짝수라 하자. 그러면, $\lambda_i = \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1} a_t \cos\theta$ 이고, $\upsilon_i = \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1} c_t \cos\theta$ 이다. 그런데, 가정에 의해 $X \in [L,U]$ 이므로, 각 t에 대해서 $l_t \le x_t \le u_t$가 성립한다. 따라서, $\cos\theta$가 양수라면 $l_t\cos\theta \le x_t\cos\theta \le u_t\cos\theta$가 성립하고, $\cos\theta$가 음수라면 $u_t\cos\theta \le x_t\cos\theta \le l_t\cos\theta$가 성립한다. 그러므로, 모든 t에 대해서 $\cos\theta$가 양수인 경우 $l_t\cos\theta$를, 음수인 경우 $u_t\cos\theta$를 합하여 구한 $\sum_{t=0}^{n-1} a_t\cos\theta$는 $\sum_{t=0}^{n-1} x_t\cos\theta$ 이하가 된다. 따라서, $\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1} a_t\cos\theta$ (= λ_i)는 $\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1} x_t\cos\theta$ (= x_i^{mbrDFT}) 이하가 된다. 또한, 모든 t에 대해서 $\cos\theta$가 양수인 경우 $u_t\cos\theta$를, 음수인 경우 $l_t\cos\theta$를 합하여 구한 $\sum_{t=0}^{n-1} c_t\cos\theta$는 $\sum_{t=0}^{n-1} x_t\cos\theta$ 이상이 된다. 따라서, $\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1} c_t\cos\theta$ (= υ_i)는 $\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1} x_t\cos\theta$ (= x_i^{mbrDFT}) 이상이 된다. 결국, i가 짝수인 모든 경우에 대해서, $\lambda_i \le x_i^{mbrDFT} \le \upsilon_i$가 성립한다. 다음으로, i가 홀수인 경우도 짝수인 경우와 동일한 방법으로 $\lambda_i \le x_i^{mbrDFT} \le \upsilon_i$ 임을 보일 수 있다. 따라서, 모든 i에 대해서 $\lambda_i \le x_i^{mbrDFT} \le \upsilon_i$가 성립하고, 정의 1에 따라 mbrDFT는 MBR-safe하다. □

다음 예제 2는 mbrDFT가 MBR-safe함을 나타내는 예이다.

예제 2. 예제 1에서와 같이 시퀀스 X가 {3.00, 2.50, 3.50, 3.00}로, MBR $[L,U]$가 L = {2.00, 1.00, 3.00, 2.00}, U = {4.00, 3.00, 5.00, 4.00}으로 주어졌다고 하자. 그리고, mbrDFT를 통해 4차원을 2차원으로 변환하려 한다. 그러면, mbrDFT의 정의에 의해, 시퀀스 X가 변환된 X^{mbrDFT}는 {6.00, -0.25}가 되고, MBR $[L,U]$가 변환된 $[L,U]^{mbrDFT}$ (=$[\Lambda,\Upsilon]$)는 Λ = {4.00, -1.50}, Υ = {8.00, 0.50}의 MBR이 된다. 이때, $4.00 \le 6.00 \le 8.00$과 $-1.50 \le -0.25 \le 0.50$이 모두 성립한다. 따라서, X와 $[L,U]$가 변환된 X^{mbrDFT}와 $[L,U]^{mbrDFT}$ 사이에는 $X^{mbrDFT} \in [L,U]^{mbrDFT}$의 관계가 성립한다. 다시 말해서, mbrDFT는 MBR-safe한 변환이다. □

제안한 mbrDFT는 DFT를 기반으로 고차원 MBR을 저차원 MBR로 직접 변환하는 최적의 MBR-safe 변환이다. 다음 따름정리는 mbrDFT가 DFT 기반의 최적 MBR-safe 변환임을 보여준다.

따름정리 1. 주어진 n-차원 MBR을 $[L,U]$라 하고, 이를 mbrDFT로 변환한 m-차원 MBR을 $[\Lambda,\Upsilon]$이라 하자. 그리고, DFT 기반의 또 다른 MBR-safe 변환이 있어, MBR $[L,U]$를 m-차원 MBR $[A,B]$로 변환한다고 하자. 그러면, $[\Lambda,\Upsilon]$은 반드시 $[A,B]$에 포함($[\Lambda,\Upsilon] \subseteq [A,B]$)된다.

증명. 부록 A 참조. □

따름정리 1의 의미는 mbrDFT가 변환한 저차원 MBR보다 더 작은 MBR을 생성하는 DFT 기반의 다른 변환은 존재하지 않음을 의미한다. 따라서, mbrDFT는 DFT를 기반으로 고차원 MBR을 저차원 MBR로 MBR-safe 변환하는 최적의 방법이라 할 수 있다.

4.2 mbrDCT: MBR-Safe DCT Transform

본 절에서는 DCT에 대한 MBR-safe 변환을 제시한다. DCT는 주로 이미지 등의 멀티미디어 데이타 압축에 많이 사용되었으나[22,23], 시계열 데이타 및 스트리밍 시계열 등의 시퀀스에 대해서도 저차원 변환 기법으로 사용되고 있다[22,24]. 차원이 n인 시퀀스 X를 DCT로 변환한 시퀀스 Y는 다음 공식 (6)에 의해 구해진 시퀀스 $\{y_0, y_1, ..., y_{n-1}\}$로 정의된다[27].

$$y_i = \frac{2 \cdot c(i)}{n}\sum_{t=0}^{n-1} x_t \cos\left(\frac{(2t+1)i\pi}{2n}\right),\ 0 \le i \le n-1,$$

$$\text{where } c(i) = \begin{cases} \sqrt{2}/2, & i = 0; \\ 1, & i = 1,2,...,n-1. \end{cases} \quad (6)$$

유사 시퀀스 매칭에서는 이와 같이 DCT로 변환된 시퀀스에서 에너지가 집중된 상위 몇 개의 차원만을 선택하는 저차원 변환을 사용한다. 다음 정의 4는 DCT 기반의 저차원 변환을 나타낸다.

정의 4. 차원이 n인 시퀀스 X를 차원 m($\le n$)으로 **DCT 저차원 변환한 시퀀스** X^{DCT}는 다음 공식 (7)에 의해 구해진 시퀀스 $\{x_0^{DCT}, x_1^{DCT}, ..., x_{m-1}^{DCT}\}$로 정의한다. 그리고, 차원이 n인 MBR $[L,U]$를 차원 m으로 **DCT 저차원 변환한 MBR** $[L,U]^{DCT}$는 시퀀스 L과 U를 각각 L^{DCT}과 U^{DCT}로 변환하여 얻은 MBR $[L^{DCT}, U^{DCT}]$로 정의한다.

$$x_i^{DCT} = \frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} x_t \cos\theta,$$

$$\text{where } c(i) = \begin{cases} \sqrt{2}/2, & i = 0; \\ 1, & i = 1,2,\ldots,n-1; \end{cases}, \quad \theta = \frac{(2t+1)i\pi}{2n},$$

$$\text{and } 0 \le i \le m-1. \tag{7}$$

그러나, 정의 4의 DCT 저차원 변환은 MBR-safe하지 못하다. 다음 예제 3은 DCT 저차원 변환이 MBR-safe하지 못하는 반례를 보여준다.

예제 3. 예제 1에서와 같이 시퀀스 X가 {3.00, 2.50, 3.50, 3.00}로, MBR $[L,U]$가 L = {2.00, 1.00, 3.00, 2.00}, U = {4.00, 3.00, 5.00, 4.00}으로 주어졌다고 하자. 그리고, DCT 저차원 변환을 통해 4차원을 2차원으로 변환하려 한다. 그러면, DCT 저차원 변환의 정의에 의해, 시퀀스 X가 변환된 X^{DCT}는 {4.24, -0.19}가 되고, MBR $[L,U]$가 변환된 $[L,U]^{DCT}$는 L^{DCT} = {2.83, -0.38}, U^{DCT} = {5.66, -0.38}의 MBR이 된다. 여기에서, $2.83 \le 4.24 \le 5.66$이므로 $l_0^{DCT} \le x_0^{DCT} \le u_0^{DCT}$가 성립한다. 그러나, $-0.38 \le -0.19 \not\le -0.38$이므로 $l_1^{DCT} \le x_1^{DCT} \le u_1^{DCT}$는 성립하지 않는다. 따라서, X와 $[L,U]$가 DCT 저차원 변환된 X^{DCT}와 $[L,U]^{DCT}$ 사이에는 $X^{DCT} \in [L,U]^{DCT}$의 관계가 성립하지 않는다. 다시 말해서, DCT 저차원 변환은 MBR-safe하지 않다. □

이와 같이 DCT 저차원 변환은 MBR-safe하지 않으므로, 본 논문에서는 기존 DCT 정의를 확장하여 다음과 같이 mbrDCT를 정의한다.

정의 5. 차원이 n인 MBR $[L,U]$를 차원 m($\le n$)으로 **mbrDCT 변환한 MBR** $[L,U]^{mbrDCT}$는 다음 공식 (8)에 의해 구해진 MBR $[\Lambda,\Upsilon]$으로 정의한다. 그리고, 시퀀스 X를 **mbrDCT 변환한 시퀀스** X^{mbrDCT}는 X를 DCT 저차원 변환한 시퀀스 X^{DCT}와 동일하다.

$$\lambda_i = \frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} a_t \cos\theta, \quad \upsilon_i = \frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} b_t \cos\theta,$$

$$\text{where } \begin{cases} a_t = l_t, b_t = u_t, & \cos\theta \ge 0; \\ a_t = u_t, b_t = l_t, & \cos\theta < 0; \end{cases}, \quad \theta = \frac{(2t+1)i\pi}{2n}, \text{ and}$$

$$0 \le i \le m-1. \tag{8}$$

정의 3의 mbrDFT와 마찬가지로, 정의 5의 mbrDCT는 DCT의 MBR-safe를 보장하기 위하여, 변환된 MBR $[\Lambda,\Upsilon]$이 원 MBR $[L,U]$내의 개별 시퀀스가 저차원 변환된 모든 점들을 포함하도록 유도된 것이다. 다음 정리 2는 mbrDCT가 MBR-safe함을 나타낸다.

정리 2. 시퀀스 X와 MBR $[L,U]$가 주어졌을 때, $X \in [L,U]$이면, X와 $[L,U]$를 각각 mbrDCT 변환한 X^{mbrDCT}와 $[L,U]^{mbrDCT}$(=$[\Lambda,\Upsilon]$) 사이에는 $X^{mbrDCT} \in [L,U]^{mbrDCT}$의 관계가 성립한다. 즉, mbrDCT는 MBR-safe하다.

증명. 모든 i에 대해서 $\lambda_i \le x_i^{mbrDCT} \le \upsilon_i$가 성립하면, $X^{mbrDCT} \in [\Lambda,\Upsilon] = [L,U]^{mbrDCT}$가 성립한다. 정의 5에 의해, $\lambda_i = \frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} a_t \cos\theta$이고, $\upsilon_i = \frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} b_t \cos\theta$이다. 그런데, 가정에 의해 $X \in [L,U]$이므로, 각 t에 대해서 $l_t \le x_t \le u_t$가 성립한다. 따라서, $\cos\theta$가 양수라면 $l_t \cos\theta \le x_t \cos\theta \le u_t \cos\theta$가 성립하고, $\cos\theta$가 음수라면 $u_t \cos\theta \le x_t \cos\theta \le l_t \cos\theta$가 성립한다. 그러므로, 모든 t에 대해서 $\cos\theta$가 양수인 경우 $l_t \cos\theta$를, 음수인 경우 $u_t \cos\theta$를 합하여 구한 $\sum_{t=0}^{n-1} a_t \cos\theta$는 $\sum_{t=0}^{n-1} x_t \cos\theta$ 이하가 된다. 따라서, $\frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} a_t \cos\theta$(= λ_i)는 $\frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} x_t \cos\theta$(= x_i^{mbrDCT}) 이하가 된다. 또한, 모든 t에 대해서 $\cos\theta$가 양수인 경우 $u_t \cos\theta$를, 음수인 경우 $l_t \cos\theta$를 합하여 구한 $\sum_{t=0}^{n-1} b_t \cos\theta$는 $\sum_{t=0}^{n-1} x_t \cos\theta$ 이상이 된다. 따라서, $\frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} b_t \cos\theta$ (= υ_i)는 $\frac{2 \cdot c(i)}{n} \sum_{t=0}^{n-1} x_t \cos\theta$ (= x_i^{mbrDCT}) 이상이 된다. 결국, 모든 i에 대해서, $\lambda_i \le x_i^{mbrDCT} \le \upsilon_i$가 성립하고, $X^{mbrDCT} \in [\Lambda,\Upsilon]$이 성립한다. 따라서, 정의 1에 따라 mbrDCT는 MBR-safe하다. □

다음 예제 4는 mbrDCT가 MBR-safe함을 나타내는 예이다.

예제 4. 예제 1에서와 같이 시퀀스 X가 {3.00, 2.50, 3.50, 3.00}로, MBR $[L,U]$가 L = {2.00, 1.00, 3.00, 2.00}, U = {4.00, 3.00, 5.00, 4.00}으로 주어졌다고 하자. 그리고, mbrDCT를 통해 4차원을 2차원으로 변환하려 한다. 그러면, mbrDCT의 정의에 의해, 시퀀스 X가 변환된 X^{mbrDCT}는 {4.24, -0.19}가 되고, MBR $[L,U]$가 변환된 $[L,U]^{mbrDCT}$(=$[\Lambda,\Upsilon]$)는 Λ = {2.83, -1.69}, Υ = {5.66, 0.92}의 MBR이 된다. 이때, $2.83 \le 4.24 \le 5.66$과 $-1.69 \le -0.19 \le 0.92$가 모두 성립한다. 따라서, X와

$[L,U]$가 변환된 X^{mbrDCT}와 $[L,U]^{mbrDCT}$ 사이에는 $X^{mbrDCT} \in [L,U]^{mbrDCT}$의 관계가 성립한다. 다시 말해서, mbrDCT는 MBR-safe한 변환이다. □

제4.1절의 mbrDFT와 마찬가지로, mbrDCT는 DCT를 기반으로 고차원 MBR을 저차원 MBR로 직접 변환하는 최적의 MBR-safe 변환이다. 즉, mbrDCT가 변환한 저차원 MBR보다 더 작은 MBR을 생성하는 DCT 기반의 다른 변환은 존재하지 않는다. 다음 따름정리는 mbrDCT가 DCT 기반의 최적 MBR-safe 변환임을 나타낸다.

따름정리 2. 주어진 n-차원 MBR을 $[L,U]$라 하고, 이를 mbrDCT로 변환한 m-차원 MBR을 $[\Lambda,\Upsilon]$이라 하자. 그리고, DCT 기반의 또 다른 MBR-safe 변환이 있어, MBR $[L,U]$를 m-차원 MBR [A,B]로 변환한다고 하자. 그러면, $[\Lambda,\Upsilon]$은 반드시 [A,B]에 포함($[\Lambda,\Upsilon] \subseteq [A,B]$)된다.

증명. 부록 B 참조. □

5. 계산 복잡도 분석

본 장에서는 저차원 MBR을 구성하는 방법들에 대한 계산 복잡도를 분석한다. 먼저, 기존 방법인 DFT 저차원 변환 후 MBR을 구성하는 방법(간략히 **DFT 복잡도**라 한다)과 mbrDFT를 사용하여 MBR을 구성하는 방법(간략히 **mbrDFT 복잡도**라 한다)를 분석하고 비교한다. 다음으로, 유사한 방법으로 **DCT 복잡도**와 **mbrDCT 복잡도**를 분석하고 비교한다. 본 논문에서는 설명의 편의상, 하나의 MBR을 변환하는데 필요한 계산 복잡도를 분석한다.

DFT 복잡도는 시퀀스 길이 n과 MBR 내에 포함되는 시퀀스 개수 m에 의해 결정된다. 즉, 하나의 시퀀스에 대한 DFT의 계산 복잡도가 $O(f(n))$이라면, m개의 시퀀스를 DFT 저차원 변환하기 위한 계산 복잡도는 $O(mf(n))$이 된다. 그런데, DFT 복잡도는 $O(n\log n)$으로 알려져 있다[27]. 따라서, DFT 복잡도는 다음 공식 (9)와 같이 $O(mn\log n)$으로 계산된다.

m개의 시퀀스×한 시퀀스의 계산 복잡도

$$= O(m)\cdot O(n\log n) = O(mn\log n) \quad (9)$$

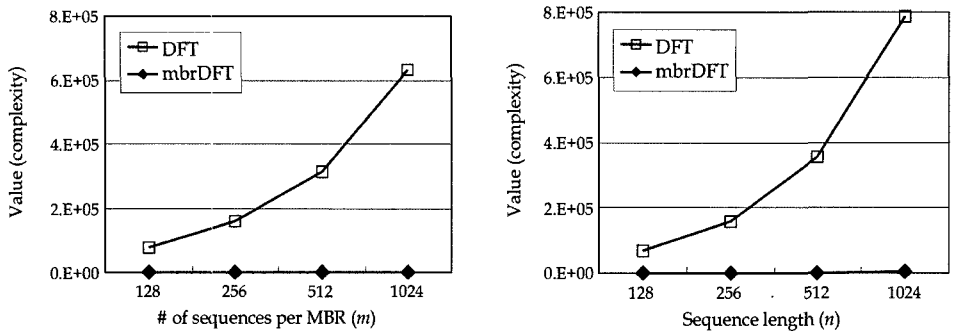
다음으로, mbrDFT의 경우는 MBR $[L,U]$를 MBR $[\Lambda,\Upsilon]\left(=[L,U]^{mbrDFT}\right)$으로 변환하기 위하여, 두 시퀀스 Λ와 Υ을 위해 각각 한번씩 총 두 번의 DFT를 수행하는 것으로 볼 수 있다. 그리고, mbrDFT에서 두 시퀀스에 대한 계산 복잡도는 DFT와 마찬가지로 각각 $O(n\log n)$이다. 따라서, 다음 공식 (10)에 의해 mbrDFT 복잡도는 $O(n\log n)$이 된다.

두 시퀀스×한 시퀀스의 계산 복잡도

$$= O(1)\cdot O(n\log n) = O(n\log n) \quad (10)$$

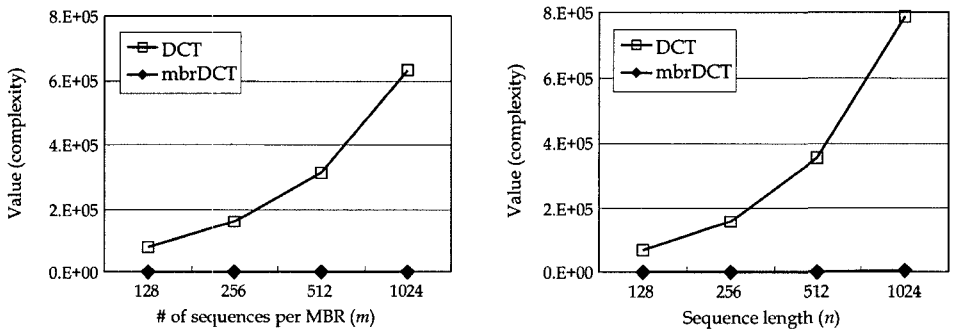
이와 같이, DFT 복잡도는 $O(mn\log n)$, mbrDFT 복잡도는 $O(n\log n)$으로 각각 구할 수 있다. 결국, 고차원 MBR 내에 포함되는 시퀀스의 개수가 많아질수록, 즉 m이 커질수록 DFT 저차원 변환과 mbrDFT의 수행 시간은 큰 차이를 나타낼 것이다. 그런데, 유사 시퀀스 매칭에서는 시퀀스(윈도우)의 길이 n이 수십~수백인 반면에, MBR내에 포함되는 시퀀스(점)의 개수 m은 수십~수천 이다[2,14]. 따라서, 많은 경우에 $m \geq n$이고, 결국 제안한 mbrDFT는 DFT에 비해 훨씬 효율적인 저차원 변환을 수행한다고 말할 수 있다. 그림 3(a)는 시퀀스 길이 n이 256이고, MBR 안에 포함되는 시퀀스 개수 m이 128, 256, 512, 1024으로 변할 때, DFT 복잡도와 mbrDFT 복잡도를 그래프로 나타낸 것이다. 그리고, 그림 3(b)는 MBR의 시퀀스 개수를 256으로 고정하고, 시퀀스 길이를 128, 256, 512, 1024로 변경했을 때의 그래프이다. 그림에서 보듯이, mbrDFT는 DFT에 비해 저차원 변환의 복잡도가 훨씬 낮음을 알 수 있다. 또한, m 혹은 n이 커질수록 그 차이가 커짐을 알 수 있다. 이는 제안한 mbrDFT가 하나의 MBR에 많은 시퀀스를 포함하거나, 시퀀스 길이가 긴 대용량 데이타 환경에 매우 적합함을 의미한다. 이러한 계산 복잡도에 의한 실제 성능 향상은 제6장의 성능 평가에서 확인할 수 있다.

다음으로, DCT 복잡도와 mbrDCT 복잡도도 유사한 방법으로 구할 수 있다. DCT 복잡도는 DFT 복잡도와 같이, $O(n\log n)$이다[27]. 따라서, DCT 복잡도는 $O(mn\log n)$, mbrDCT 복잡도는 $O(n\log n)$으로 구할 수 있다. 그림 4는 DCT와 mbrDCT의 계산 복잡도를 그래프로 그린 것이다. 그런데, 그림 4의 그래프는 그림 3의 DFT와 mbrDFT의 그래프와 동일함을 알 수 있다. 이는 DCT 복잡도와 DFT 복잡도는 $O(mn\log n)$으로 동일하고, mbrDCT 복잡도와 mbrDFT 복잡도 역시 $O(n\log n)$으로 동일하기 때문이다. 그림 4의 그래프를 보면, 그림 3의 DFT 및 mbrDFT의 경우와 같이, mbrDCT의 계산 복잡도가 DCT 복잡도에 비해 훨씬 낮음을 알 수 있다. 마찬가지로, 이에 따른 성능 향상은 제6장의 성능 평가에서 확인할 수 있다.



(a) Complexity when varying m (b) Complexity when varying n

그림 3 DFT와 mbrDFT의 계산 복잡도 그래프



(a) Complexity when varying m (a) Complexity when varying n

그림 4 DCT와 mbrDCT의 계산 복잡도 그래프

6. 성능 평가

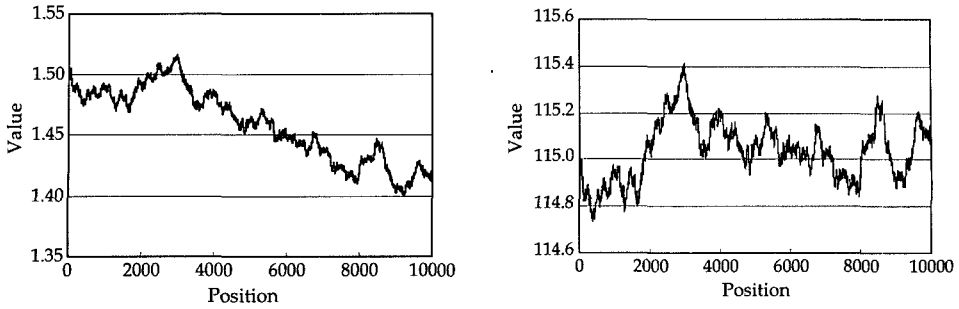
본 장에서는 실제 실험을 통해 mbrDFT와 DFT, 그리고 DCT와 mbrDCT의 성능을 비교한다. 제6.1절에서는 실험 데이타와 실험 환경을 소개하고, 제6.2절에서는 실험 결과를 설명한다.

6.1 실험 데이타 및 실험 환경

제안한 방법의 우수성을 입증하기 위하여 두 가지 종류의 합성 데이타(synthetic data)를 사용하였다. 첫 번째 데이타는 데이타 시퀀스의 시작 엔트리를 1.5로 하고, 각 엔트리에 (-0.001, 0.001) 사이의 임의의 값 하나를 더하여 다음 엔트리를 구하는 방식으로 생성된 100만개의 랜덤 워크 시리즈(random walk series)이다. 이 데이타는 기존 유사 시퀀스 매칭 연구[2,5,14]에서 사용한 것으로서, 이를 *WALK-DATA*라 한다. 그림 5(a)는 WALK-DATA 일부를 나타낸다. 두 번째 데이타는 스트리밍 시계열로서, 참고문헌 [15,16]에서와 같이 랜덤 워크 시리즈 y_i에 대한 함수로서 $x_i = 100 \cdot (\sin(0.1 \cdot y_i) + 1 + i/1000000)$을 사용하여 생성한 100만개의 데이타이다. 여기에서, 랜덤 워크 시리즈인 y_i로는 첫 번째 데이타인 WALK-DATA를 사용하였으며, 이 데이타를 *SINE-DATA*라 한다. 그림 5(b)는 SINE-DATA 일부를 나타낸다.

고차원 MBR 구성은 서브시퀀스 매칭에서 사용하는 전체 데이타를 여러 개의 작은 시퀀스(윈도우)[2,14]로 나누는 방법을 사용하였다. 이때, 시퀀스 길이로는 128, 256, 512, 1024를 사용하였고, 하나의 MBR에 포함되는 시퀀스 개수 역시 128, 256, 512, 1024를 사용하였다. 그리고, 참고문헌 [1]에서와 같이 각 시퀀스는 1~4-차원의 시퀀스(점)로 저차원 변환되는 것으로 실험하였다. 즉, 저차원 변환을 통해 추출하는 특성의 개수를 1~4로 하여 실험하였다[1]. 실험 방법으로는 DFT 관련 두 가지, DCT 관련 두 가지 등 총 네 가지 방법을 실험하였다. 실험 결과에서는 DFT를 사용해 개별 시퀀스를 저차원 변환한 후 MBR을 구성하는 기존 DFT 변환 방법을 *DFT*, 기존 DCT 변환 방법을 *DCT*라 간략히 나타낸다.

실험을 수행한 하드웨어 플랫폼은 Intel Pentium IV 2.80 GHz CPU, 512 MB RAM, 70.0GB 하드디스크를 장착한 PC이며, 소프트웨어 플랫폼은 GNU/Linux



(a) WALK-DATA (b) SINE-DATA

그림 5 실험 데이타의 일부분($=\{x_0, x_1, ..., x_{10000}\}$).

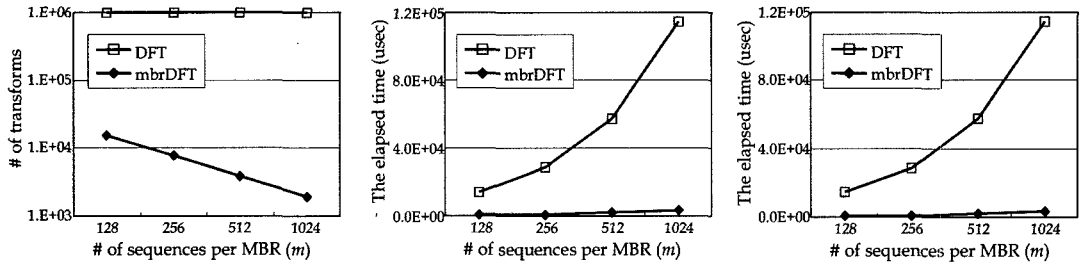
Version 2.6.6 운영 체제이다. 실험 결과로는 각 방법의 변환 횟수, 실제 수행 시간을 성능 요소로서 측정하였다. 또한, 각 방법에 의해 저차원 변환된 MBR들을 비교함으로써, 제안한 mbrDFT와 mbrDCT이 실용적으로 사용 가능한 변환 방법임을 보인다.

6.2 실험 결과

본 절에서는 네 가지 저차원 변환에 대한 성능 평가 결과를 설명한다. 먼저, 실험 1)에서는 시퀀스 길이(n)를 고정하고 MBR에 포함되는 시퀀스 개수(m)를 달리하면서 변환 횟수와 실제 수행 시간을 측정하였다. 그리고, 실험 2)에서는 MBR에 포함되는 시퀀스 개수를 고정하고, 시퀀스 길이를 달리하면서 실험을 수행하였다. 마지막으로, 실험 3)에서는 변환된 MBR의 차원 길이에 대해서 DFT와 mbrDFT, 그리고 DCT와 mbrDCT를 각각 비교하였다.

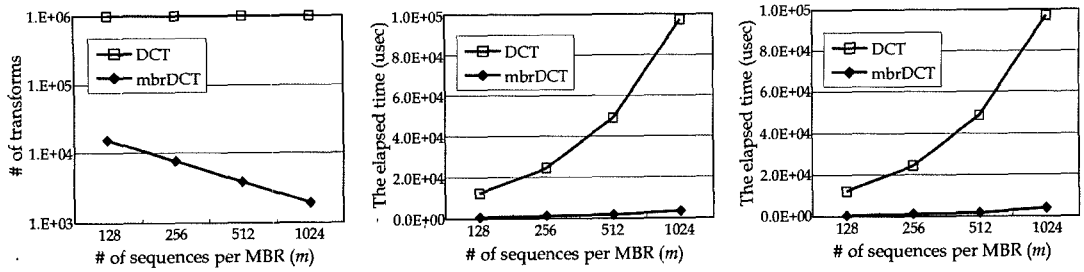
실험 1) 시퀀스 길이를 고정하고 MBR의 시퀀스 개수를 달리한 경우의 성능 평가

그림 6은 DFT와 mbrDFT에 대해 시퀀스 길이 n을 256으로 고정하고 MBR에 포함되는 시퀀스 개수 m을 128, 256, 512, 1024로 달리하면서, 저차원 변환의 전체 횟수와 MBR 하나에 대한 평균 수행 시간을 측정한 결과이다. 이때, 추출하는 차원의 개수는 참고문헌 [1]에서와 같이 두 개로 하였다. 그림 6(a)는 WALK-DATA 및 SINE-DATA에 대한 저차원 변환 횟수를 나타내며, 그림 6(b)와 6(c)는 각각 WALK-DATA와 SINE-DATA에 대한 실제 수행 시간을 나타낸다. 그림 6(a)를 보면, 제안한 mbrDFT가 DFT에 비해 변환 횟수를 획기적으로 줄였음을 알 수 있다(세로축이 로그 눈금에 유의한다). 이는 DFT가 MBR 내의 모든 시퀀스에 대해서 저차원 변환을 수행하는 반면에, mbrDFT는 각 MBR에서 단 두 번의 변환만을 수행하기 때문이다. 특히, m이 커질수록 mbrDFT의 변환 횟수가 줄어드는데, 이는 m이 커지면 mbrDFT가 변환 대상으로 하는 MBR 개수가 줄어들기 때문이다. 이와 같이 변환 횟수를 줄임으로써, 그림 6(b) 및 6(c)에서 보듯이 저차원 변환을 위한 수행 시간이 크게 줄었음을 알 수 있다. 또한, 제5장의 그림 3(a)에서 분석한 바와 같이, MBR에 포함된 시퀀스 개수가 많을수록 성능 차이가 커짐을 확인할 수 있다. 그리고, 그림 6(b)의 WALK-DATA 결과와 그림 6(c)의 SINE-DATA 결과는 유사함을 알 수 있다. 이는 DFT와 mbrDFT 모두 저차원 변환에 걸리는 시간은 데이타의 종류와는 관계가 없기 때문이다. 실험 결과를 요약하면, mbrDFT는 DFT에 비해 변환 횟수를 평균 $1/136$로 줄이고, 성능을 평균 31배 향상시킨 것으로 나타났다. 그림 6의 실험 결과에서, DFT 변환



(a) Number of transforms (b) The elapsed time(WALK-DATA) (c) The elapsed time(SINE-DATA)

그림 6 MBR의 시퀀스 개수(m)를 달리한 경우 DFT와 mbrDFT의 저차원 변환 횟수와 수행 시간



(a) Number of transforms (b) The elapsed time(WALK-DATA) (c) The elapsed time(SINE-DATA)

그림 7 MBR의 시퀀스 개수(m)를 달리한 경우 DCT와 mbrDCT의 저차원 변환 횟수와 수행 시간

횟수와 실제 수행 시간에 차이가 있는 이유는 저차원 변환을 위한 수행 시간 이외에도 시퀀스에 대한 I/O, MBR 구성 등에 추가적인 수행 시간이 소요되기 때문이다.

그림 7은 MBR에 포함되는 시퀀스 개수(m)를 달리하면서 DCT와 mbrDCT에 대한 저차원 변환 횟수와 실제 수행 시간을 측정한 결과이다. 그림 6에서와 마찬가지로, 그림 7(a)는 저차원 변환 횟수를 나타내고, 그림 7(b)와 7(c)는 각각 WALK-DATA와 SINE-DATA에 대한 두 방법의 실제 수행시간을 각각 나타낸다. 그림 7의 실험 결과의 결과를 보면, 그림 6의 mbrDFT와 DFT에서와 같이, mbrDCT가 DCT에 비해 성능을 크게(평균 26배) 향상시킨 것으로 나타났다.

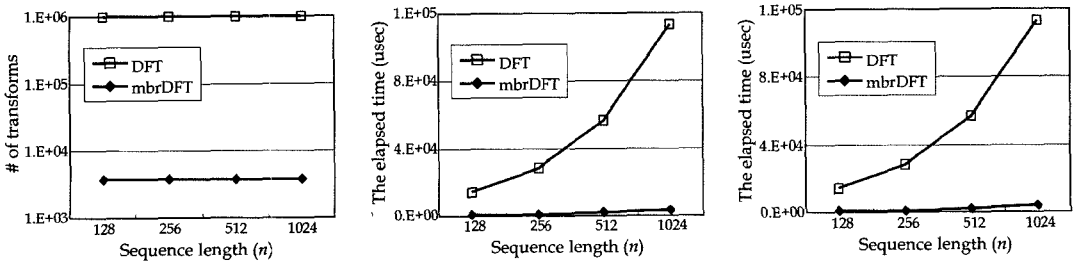
실험 2) MBR의 시퀀스 개수를 고정하고 시퀀스 길이를 달리한 경우의 성능 평가

그림 8은 DFT와 mbrDFT에 대해 MBR에 포함되는 시퀀스 개수 m을 256으로 고정하고 시퀀스 길이 n을 128, 256, 512, 1024로 달리하면서, 저차원 변환의 전체 횟수와 MBR 하나에 대한 평균 수행 시간을 측정한 결과이다. 그림 8(a)는 WALK-DATA 및 SINE-DATA에 대한 저차원 변환 횟수를 나타내며, 그림 8(b)와 8(c)는 각각 WALK-DATA와 SINE-DATA에 대한 실제 수행 시간을 나타낸다. 그림 8(a)를 보면, 저차원 변환 횟수는 시퀀스 길이와 관계 없이 일정함을 알 수 있다. 이는 저차원 변환 횟수가 DFT의 경우 시퀀스 개수, mbrDFT의 경우 MBR 개수에만 의존적일 뿐 시퀀스 길이와는 무관하기 때문이다. 그림 8(b)와 8(c)의 수행 시간을 보면, mbrDFT는 DFT에 비해 수행 시간을 크게 줄였음을 알 수 있다. 특히, 제5장의 그림 3(b)에서 분석한 바와 같이, 시퀀스 길이가 커질수록 성능 차이가 커짐을 확인할 수 있다. 그림 8의 실험 결과, mbrDFT는 DFT에 비해 성능을 평균 30배 향상 시킨 것으로 나타났다.

그림 9는 시퀀스 길이(n)를 달리하면서 DCT와 mbrDCT에 대한 저차원 변환 횟수와 실제 수행 시간을 측정한 결과이다. 그림 8에서와 마찬가지로, 그림 9(a)는 DCT와 mbrDCT의 저차원 변환 횟수를 나타내고, 그림 9(b)와 9(c)는 각각 WALK-DATA와 SINE-DATA에 대한 두 방법의 실제 수행시간을 각각 나타낸다. 그림 9의 DCT와 mbrDCT의 실험 결과는 그림 8의 mbrDFT와 DFT의 결과와 매우 유사함을 알 수 있다. 그림 9의 실험 결과를 요약하면, mbrDCT가 DCT에 비해 성능을 크게(평균 26배) 향상시킨 것으로 나타났다.

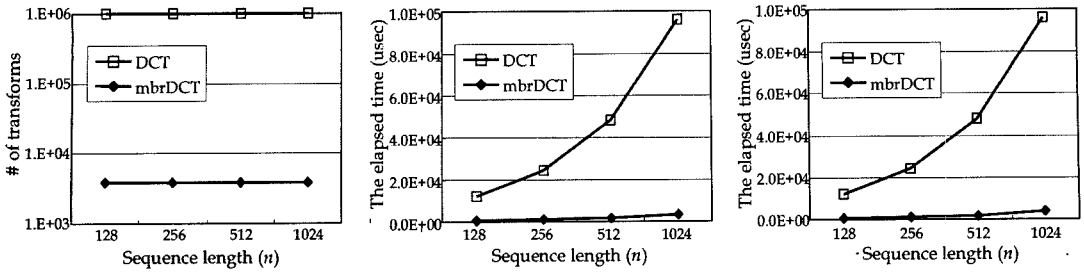
실험 3) 변환된 MBR의 차원 길이 비교

본 실험에서는 기존 변환에 의한 MBR과 제안한



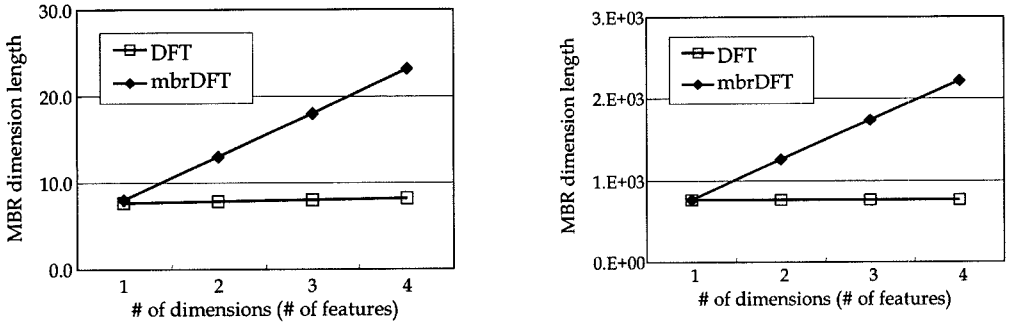
(a) Number of transforms (b) The elapsed time(WALK-DATA) (c) The elapsed time(SINE-DATA)

그림 8 시퀀스 길이(n)를 달리한 경우 DFT와 mbrDFT의 저차원 변환 횟수와 수행 시간



(a) Number of transforms (b) The elapsed time(WALK-DATA) (c) The elapsed time(SINE-DATA)

그림 9 시퀀스 길이(n)를 달리한 경우 DCT와 mbrDCT의 저차원 변환 횟수와 수행 시간



(a) WALK-DATA (b) SINE-DATA

그림 10 추출하는 차원 개수를 달리한 경우의 DFT와 mbrDFT의 MBR 차원 길이

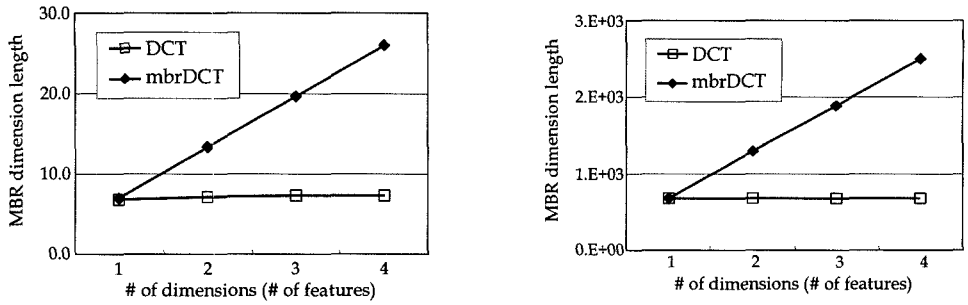
MBR-safe 변환에 의한 MBR의 차원 길이를 비교한다. 여기에서 **MBR의 차원 길이**란 저차원 변환된 MBR을 이루는 각 차원의 길이를 합한 값으로 정의한다.[2] 이 실험을 수행한 이유는 제안한 mbrDFT와 mbrDCT가 고차원 MBR의 모든 시퀀스를 고려한 반면에, 실제 유사 시퀀스 매칭에서는 고차원 MBR내에 유한 개의 시퀀스만이 포함되기 때문이다. 즉, 이론적으로는 제안한 mbrDFT와 mbrDCT가 고차원 MBR 자체에 대한 최적의 저차원 MBR을 구성하나, 실제로 고차원 MBR은 수십~수천 개의 유한한 시퀀스만을 대표하기 때문이다. 이러한 이유에 의해 DFT(혹은 DCT)에 의한 MBR과 mbrDFT(혹은 mbrDCT)에 의한 MBR은 동일하지 않게 된다. 따라서, 본 실험에서는 이들 두 MBR을 정량적으로 비교하기 위하여, 각 MBR의 차원 길이를 사용한다.

그림 10은 DFT와 mbrDFT에 대해 MBR 차원 길이를 비교한 것이다. 이때, 시퀀스 길이(n)와 MBR에 포함된 시퀀스 개수(m)를 각각 256으로 설정하고, 추출하는 차원의 수를 1, 2, 3, 4로 변경하며 실험하였다. 그림 10(a)는 WALK-DATA에 대한 실험 결과이고, 그림 10(b)는 SINE-DATA에 대한 실험 결과이다. 그림을 보면, 차원의 수가 2 이상인 경우 mbrDFT가 DFT에 비해 MBR 차원 길이가 커짐을 알 수 있다. 이는 mbrDFT의 경우 고차원 MBR내의 모든 가능한 무한 개 시퀀스를 고려하는 반면에, DFT는 실제의 유한 개 시퀀스만을 고려하기 때문이다.[3] 반면에, 차원 개수가 1인 경우에는 DFT와 mbrDFT의 MBR 차원 길이에 큰 차이가 없다(0.2%~2.6%). 참고문헌 [1]에서 지적한 바와 같이, 차원의 개수가 1~2인 경우와 3 이상인 경우의 특성 추출 효과는 큰 차이가 없으므로, 제안한 mbrDFT는 한 개 혹은 두 개 차원을 사용할 경우 매우 유리한 방법이라 할 수 있다.

다음으로, 그림 11은 DCT와 mbrDCT에 대해 MBR 차원 길이를 비교한 것이다. 그림을 보면, DCT와 mbrDCT에 대한 MBR 차원 길이 그래프는 그림 10의 DFT와 mbrDCT에 대한 그래프와 그 경향이 거의 유사함을 알 수 있다. 즉, 차원의 수가 1인 경우 mbrDCT

2) MBR 비교를 위해 각 차원의 길이를 합한 차원 길이를 사용한 이유는, 유사 시퀀스 매칭에서 다차원 색인에 대한 범위 질의는 특정 점을 중심으로 정 사각형 모양을 가지는데, 이때 MBR의 각 차원 길이가 길수록 검색 결과로 선택될 가능성이 높기 때문이다.

3) 실제 실험 결과, 고차원 MBR내에 매우 다양하고 많은 시퀀스(수천 개 이상)가 포함될 경우 DFT와 mbrDFT의 MBR 차원 길이 차이는 크게 줄어드는 것으로 나타났다.



(a) WALK-DATA (b) SINE-DATA

그림 11 추출하는 차원 개수를 달리한 경우의 DCT와 mbrDCT의 MBR 차원 길이

와 DCT의 차원 길이에 차이가 거의 없으나(0.2%~2.9%), 2 이상인 경우 mbrDCT가 DCT에 비해 MBR 차원 길이가 커짐을 알 수 있다. 이는 DFT 및 mbrDFT와 동일한 결과로서, MBR 차원 길이의 차이는 저차원 변환의 종류에 크게 관계 없음을 의미한다.

7. 결 론

본 논문에서는 고차원 MBR을 저차원 MBR로 변환하는 정형적인 방법을 제안하였다. 유사 시퀀스 매칭에서는 다차원 색인의 고차원 문제를 피하고, 색인의 저장 공간을 줄이기 위하여 저차원 변환을 사용한다. 본 논문에서는 기존 유사 시퀀스 매칭에서 많은 수의 고차원 시퀀스를 저차원 변환한 후 이들 변환된 저차원 시퀀스들을 포함하는 저차원 MBR을 구성하는데 주목하였다. 그 결과, 고차원 MBR 자체를 직접 저차원 MBR로 변환하는 개념을 제시하고, 이를 통하여 유사 시퀀스 매칭에서 필요한 저차원 변환 횟수를 획기적으로 줄이는 방법을 제안하였다.

본 논문의 공헌은 다음과 같이 1) MBR-safe 변환의 개념을 제시하고, 2) DFT와 DCT에 대한 MBR-safe 변환을 제안하였으며, 3) 분석과 실험을 통해 제안한 변환의 우수성을 입증한 세 가지로 요약할 수 있다.

첫째, 고차원 MBR을 저차원 MBR로 변환하기 위한 변환의 *MBR-safe* 개념을 정형적으로 제안하였다. 어떤 변환이 MBR-safe하다 함은 고차원 MBR을 직접 변환한 저차원 MBR이 개별 고차원 시퀀스가 변환된 저차원 시퀀스를 모두 포함함을 의미한다. 이러한 MBR-safe 개념을 사용하면, 개별 시퀀스를 변환하여 저차원 MBR을 구성하는 대신, 고차원 MBR 자체를 변환하여 저차원 MBR을 구성할 수 있다.

둘째, 기존 저차원 변환 중에서 가장 널리 사용되는 DFT와 DCT에 대해서 각각 MBR-safe 변환을 제안하였다. 우선, 기존 DFT와 DCT가 MBR-safe하지 않음을 보이고, DFT와 DCT를 확장한 *mbrDFT*와 *mbrDCT*를 각각 정의하였다. 또한, 이들 mbrDFT와 mbrDCT가 MBR-safe함을 정리 1과 2에서 정형적으로 증명하였다. 그리고, 제안한 mbrDFT(혹은 mbrDCT)가 고차원 MBR을 저차원 MBR로 직접 변환하는 DFT(혹은 DCT) 기반의 최적 MBR-safe 변환임을 따름정리 1과 2에서 증명하였다.

셋째, 분석과 실험을 통해 MBR-safe 변환의 우수성을 입증하였다. 우선, 기존 저차원 변환과 MBR-safe 변환의 계산 복잡도를 유도하여, 제안한 MBR-safe 변환의 우수성을 분석적으로 설명하였다. 다음으로, 실험을 통해 제안한 mbrDFT 및 mbrDCT를 사용하면 저차원 변환 횟수를 크게 줄이고 성능을 향상 시킴을 보였다.

본 논문에서 제시한 MBR-safe 개념은 유사 시퀀스 매칭뿐 아니라, 고차원 MBR을 저차원 변환해야 하는 많은 연구에 활용될 수 있다. 따라서, 향후 연구로는 MBR 개념을 사용하는 GIS 등의 여러 응용에 본 연구 결과를 적용하는 것이다. 또한, DFT와 DCT 이외에 Wavelet, PAA, SVD 등의 다양한 저차원 변환에 대해서도 MBR-safe 변환 연구가 필요하다.

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] Kim, S.-W., Yoon, J., Park, S., and Won, J.-I. "Shape-based Retrieval in Time-Series Databases," *Journal of Systems and Software*, Vol. 79, No. 2, pp. 191-203, Feb. 2006.

[4] Wu, H., Salzberg, B., and Zhang, D., "Online Event-driven Subsequence Matching Over Financial Data Streams," In *Proc. of Int'l Conf. on Management of Data*, ACM SIGMOD, Paris, France, pp. 23-34, June 2004.

[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] 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)*, Singapore, pp. 65-79, Apr. 2006.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Keogh, E. 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, California, pp. 151-162, May 2001.

[13] Yi, B.-K., Jagadish, H. V., and Faloutsos, C., "Efficient Retrieval of Similar Time Sequences Under Time Warping," In *Proc. the 14th Int'l Conf. on Data Engineering(ICDE)*, IEEE, Orlando, Florida, pp. 201-208, Feb. 1998.

[14] 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.

[15] 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.

[16] Gao, L., Yao, Z., and Wang, X. S., "Evaluating Continuous Nearest Neighbor Queries for Streaming Time Series via Pre-fetching," In *Proc. In'tl Conf. on Information and Knowledge Management*, ACM CIKM, McLean, Virginia, pp. 485-492, Nov. 2002.

[17] Rafiei, D. and Mendelzon, A. O., "Querying Time Series Data Based on Similarity," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 12, No. 5, pp. 675-693, Sept./Oct. 2000.

[18] Agrawal, R., Lin, K.-I., Sawhney, H. S., and Shim, K., "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," In *Proc. the 21st Int'l Conf. on Very Large Data Bases*, Zurich, Switzerland, pp. 490-501, Sept. 1995.

[19] Chu, K. W. and Wong, M. H., "Fast Time-Series Searching with Scaling and Shifting," In *Proc. the 15th Symposium on Principles of Database Systems*, ACM PODS, Philadelphia, Pennsylvania, pp. 237-248, June 1999.

[20] Kim, S.-W., Park, S, and Chu, W. W., "Efficient Processing of Similarity Search Under Time Warping in Sequence Databases: An Index-based Approach," *Information Systems*, Vol. 29, No. 5, pp. 405-420, July 2004.

[21] Park, S., Chu, W. W., Yoon, J., and Won, J., "Similarity Search of Time-Warped Subsequences via a Suffix Tree," *Information Systems*, Vol. 28, No. 7, pp. 867-883, Oct. 2003.

[22] Hjaltason, G. R. and Samet, H., Incremental Similarity Search in Multimedia Databases, Dept. of Computer Science, University of Maryland, College Park, Technical Report 4199, Nov. 2000.

[23] Zhao, D., Gao, W., and Chan, Y. K., "Morphological Representation of DCT Coefficients for Image Compression," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 12, No. 9, pp. 819-823, Sept. 2002.

[24] 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.

[25] Natsev, A., Rastogi, R., and Shim, K., "WALRUS: A Similarity Retrieval Algorithm for Image Databases," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 16, No. 3, pp. 301-316, Mar. 2004.

[26] Korn, F., Jagadish, H. V., and Faloutsos, C., "Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences," In *Proc. of Int'l Conf. on Management of Data*, ACM SIGMOD, Tucson, Arizona, pp. 289-300, June 1997.

[27] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T., Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, 2nd Ed., 1992.

부록 A

따름정리 1의 증명: MBR $[L,U]$를 MBR [A,B]로 변환하는 MBR-safe 변환을 T라 하자. 이때, T에 의한 [A,B]가 mbrDFT에 의한 $[\Lambda,\Upsilon]$에 완전히 포함($[A,B]\subset[\Lambda,\Upsilon]$)된다고 가정하자. 그러면, $0\le i\le m-1$인 최소한 하나의 i에 대해 $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$나 $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$가 성립해야 하는데, 이것이 모순임을 증명한다. 이 두 가지 경우와 i가 짝수 혹은 홀수인 경우를 나누어 다음 네 가지로 구분하여 증명한다.

1) $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$이고 i가 짝수인 경우: 다음 공식 (11)에 의해 구성되는 시퀀스 X가 있다고 하자.

$$X=\{x_0,x_1,\ldots,x_{n-1}\},\ \text{where } x_t=\begin{cases}l_t, & \cos\left(-2\pi\lfloor i/2\rfloor t/n\right)\ge 0\\ u_t, & \cos\left(-2\pi\lfloor i/2\rfloor t/n\right)<0\end{cases} \tag{11}$$

그러면, 시퀀스 X의 모든 x_t에 대해 $l_t\le x_t\le u_t$가 성립하므로 X는 MBR $[L,U]$내에 포함된다. 그런데, 이러한 시퀀스 X가 DFT에 의해 변환된 X^{DFT}의 x_i^{DFT}를 구하는 과정은 MBR $[\Lambda,\Upsilon]\left(=[L,U]^{mbrDFT}\right)$에서 Λ의 λ_i를 구하는 과정과 동일하다. 따라서, x_i^{DFT}와 λ_i의 값은 동일하다. 여기서, 변환 T가 MBR-safe하려면 $\lambda_i=x_i^{DFT}\ge\alpha_i$가 성립하여야 하는데, 이는 $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$라는 상기 가정에 위배된다. 따라서, $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$는 성립하지 않는다.

2) $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$이고 i가 홀수인 경우: 이 경우는 cos() 대신 sin()을 사용하는 것을 제외하고는 경우 1)과 동일하게 $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$가 성립하지 않음을 증명할 수 있다.

3) $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$이고 i가 짝수인 경우: 이 경우는 λ_i와 α_i 대신 υ_i와 β_i를 사용하여 경우 1)과 동일한 방법으로 $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$가 성립하지 않음을 증명할 수 있다.

4) $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$이고 i가 홀수인 경우: 이 경우는 λ_i와 α_i 대신 υ_i와 β_i를 사용하고, cos() 대신 sin()을 사용하여 경우 1)과 동일한 방법으로 $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$가 성립하지 않음을 증명할 수 있다.

상기 경우 1)~4)는 T에 의한 [A,B]가 mbrDFT에 의한 $[\Lambda,\Upsilon]$에 완전히 포함($[A,B]\subset[\Lambda,\Upsilon]$)된다는 가정이 잘못되었음을 의미한다. 결국, T가 DFT 기반의 MBR-safe한 변환이라면, 반드시 $[\Lambda,\Upsilon]\subseteq[A,B]$이 성립한다. □

부록 B

따름정리 2의 증명: MBR $[L,U]$를 MBR [A,B]로 변환하는 MBR-safe 변환을 T라 하자. 이때, T에 의한 [A,B]가 mbrDCT에 의한 $[\Lambda,\Upsilon]$에 완전히 포함($[A,B]\subset[\Lambda,\Upsilon]$)된다고 가정하자. 그러면, $0\le i\le m-1$인 최소한 하나의 i에 대해 $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$나 $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$가 성립해야 하는데, 이것이 모순임을 증명한다. 이 두 가지 경우를 다음과 같이 구분하여 증명한다.

1) $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$인 경우: 다음 공식 (12)에 의해 구성되는 시퀀스 X가 있다고 하자.

$$X=\{x_0,x_1,\ldots,x_{n-1}\},\ \text{where } x_t=\begin{cases}l_t, & \cos\left(\dfrac{(2t+1)i\pi}{2n}\right)\ge 0\\ u_t, & \cos\left(\dfrac{(2t+1)i\pi}{2n}\right)<0\end{cases} \tag{12}$$

그러면, 시퀀스 X의 모든 x_t에 대해 $l_t\le x_t\le u_t$가 성립하므로 X는 MBR $[L,U]$내에 포함된다. 그런데, 이러한 시퀀스 X가 DCT에 의해 변환된 X^{DCT}의 x_i^{DCT}를 구하는 과정은 MBR $[\Lambda,\Upsilon]\left(=[L,U]^{mbrDCT}\right)$에서 Λ의 λ_i를 구하는 과정과 동일하다. 따라서, x_i^{DCT}와 λ_i의 값은 동일하다. 여기서, 변환 T가 MBR-safe하려면 $\lambda_i=x_i^{DCT}\ge\alpha_i$가 성립하여야 하는데, 이는 $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$라는 상기 가정에 위배된다. 따라서, $\lambda_i<\alpha_i\le\beta_i\le\upsilon_i$는 성립하지 않는다.

2) $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$인 경우: 이 경우는 λ_i와 α_i 대신 υ_i와 β_i를 사용하여 경우 1)과 동일한 방법으로 $\lambda_i\le\alpha_i\le\beta_i<\upsilon_i$가 성립하지 않음을 증명할 수 있다.

상기 경우 1)과 2)는 T에 의한 [A,B]가 mbrDCT에 의한 $[\Lambda,\Upsilon]$에 완전히 포함($[A,B]\subset[\Lambda,\Upsilon]$)된다는 가정이 잘못되었음을 의미한다. 결국, T가 DCT 기반의 MBR-safe한 변환이라면, 반드시 $[\Lambda,\Upsilon]\subseteq[A,B]$이 성립한다.

문 양 세
정보과학회논문지 : 데이타베이스
제 33 권 제 1 호 참조