다중속성 시계열 데이타베이스의 효율적인 유사 검색

이 상 준†

요 약

시계열에 대한 색인 및 검색 연구는 하나의 속성으로 구성된 시계열에 대하여 주로 수행되어 왔다. 그러나 음악, 비디오 등의 멀티미디어 데이타베이스는 다중속성 시계열 데이타베이스에서 유사 검색을 다룰 수 있어야 한다. 기존의 다중속성 시계열 데이타베이스에 대한 연구는 두 다중속성 시퀀스간의 유사도로 속성 간의 거리의 누적을 사용하고 있기에, 개별적인 속성 시퀀스에 대한 정보를 상실하게 된다. 본 연구에서는 이러한 문제를 해결하기 위해 속성 시퀀스 측면에서 다중속성 시계열 데이타베이스의 유사검색 기법을 제안한다. 제안된 기법은 검색 공간을 효율적으로 줄일 수 있으며, 착오 누락이 없음을 보장한다. 또한 실험을 통해 제안된 기법의 성능 향상을 확인하였다.

키워드 : 유사검색, 시계열, 데이타베이스

Efficient Similarity Search in Multi-attribute Time Series Databases

Sangjun Lee†

ABSTRACT

Most of previous work on indexing and searching time series focused on the similarity matching and retrieval of one-attribute time series. However, multimedia databases such as music, video need to handle the similarity search in multi-attribute time series. The limitation of the current similarity models for multi-attribute sequences is that there is no consideration for attributes' sequences. The multi-attribute sequences are composed of several attributes' sequences. Since the users may want to find the similar patterns considering attributes's sequences, it is more appropriate to consider the similarity between two multi-attribute sequences in the viewpoint of attributes' sequences. In this paper, we propose the similarity search method based on attributes's sequences in multi-attribute time series databases. The proposed method can efficiently reduce the search space and guarantees no false dismissals. In addition, we give preliminary experimental results to show the effectiveness of the proposed method.

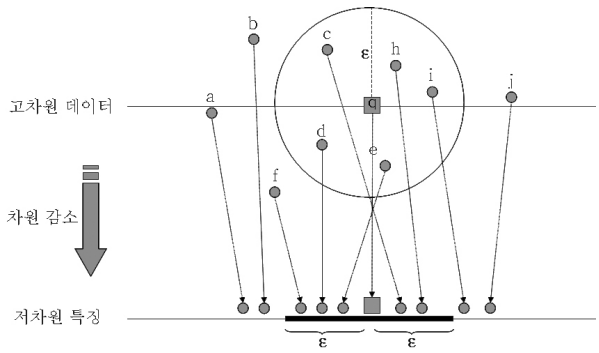
Key Words : Similarity Search, Time Series, Database

1. 서 론

시간의 흐름에 따라 순차적으로 생성되는 데이타의 연속적인 모임인 시퀀스(sequence)는 컴퓨터에 저장되는 데이타에 있어서 많은 부분을 차지하고 있다[1]. 이러한 시퀀스 중에서 숫자로 그 값이 표현되는 것을 시계열(time series)이라 한다. 시계열 데이타베이스에서 주어진 질의와 유사한 시퀀스를 검색하는 것은 멀티미디어 정보 검색과 같은 분야에서 다양하게 사용되고 있다. 시계열 데이타베이스에서 유사한 시퀀스를 검색하는데 있어 중요한 점은 검색 성능을 향상시키는 것이다. 순차 검색의 경우 질의 시퀀스와 데이타베이스 내의 전체 시퀀스 간의 유사도 계산을 수행해야하므로 데이타베이스의 크기가 커짐에 따라 성능이 떨어지게 된다. 일반적으로 검색 성능을 향상시키기 위해 R-tree[2] 또는 R^*-tree[3]와 같은 공간 접근 기법을 이용하여 효율적인 유사 검색을 수행하게 된다. 그러나 시퀀스와 같은 고차원 데이타를 그대로 공간 접근 기법을 이용하여 색인하는 방식은 차원의 저주(dimensionality curse)[1, 4]라는 현상에 의해 급격한 성능 저하가 발생한다. 이러한 문제를 해결하기 위해 시퀀스에서 차원 감소 기법(dimensionality reduction method)을 이용하여 특징을 추출하고, 색인하는 기법이 일반적으로 시계열 데이타베이스의 유사 검색에 주로 이용되고 있다.

차원 감소 기법을 이용한 시계열과 같은 고차원 데이타 색인에 있어 발생하는 문제는 착오해답(false alarm) 및 착오 누락(false dismissals)이 있다. (그림 1)은 차원 감소에 따른 착오 누락 및 착오 해답의 경우를 보여주고 있다. 실제 고차원 상에서 질의 g와 거리 차가 ε 이내인 i가 저차원 특징 공간 상에서 유사하지 않다고 판정되는 것과 같은 경우를 착오 누락이라 하며, 반대로 f와 같이 실제 거리가 ε보다 큼에도 불구하고 저차원 공간 상에서 유사하다고 판정되는 것을 착오 해답이라고 한다.

※본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.
† 정 회 원 : 숭실대학교 컴퓨터학부 조교수
논문접수 : 2007년 4월 26일, 심사완료 : 2007년 10월 24일



착오 누락(false dismissal) : i
착오 해답(false alarm) : f

(그림 1) 차원 감소에 의한 착오 누락 및 착오 해답의 예

차원 감소 기법을 이용한 시계열 데이타를 색인할 때 중요한 점은 순차 검색보다 효율적이면서 착오 누락(false dismissals)이 없다는 것을 보장해야하며, 특징 공간 상의 거리가 실제 유사도보다 작거나 같다는 하한 조건(lower bound condition)[1]을 만족해야한다. 착오 해답의 경우 후처리 과정에서 질의 시퀀스와 실제 거리를 계산하여 제거하게 된다.

시계열 데이타에 대한 유사 검색은 주로 하나의 속성이 시간에 따라 그 값이 변화하는 내용을 검색하는 데 일차원 시계열 데이타 검색에 초점을 맞추어 왔다. 그러나 음악, 비디오 등의 멀티미디어 데이타와 같이 여러 속성으로 구성되어 있고, 각 속성이 시간에 따라 변화하는 다중속성 시계열에 대한 유사 검색 연구는 미비한 실정이다. m개의 속성을 가지며 길이가 n인 다중속성 시퀀스 S는 다음과 같이 정의된다. $S_j[i]$는 j번째 속성의 i번째 요소를 나타낸다.

정의 1 (다중속성 시퀀스) m개의 속성을 가지며 길이가 n인 시퀀스 S는 다음과 같이 정의된다.

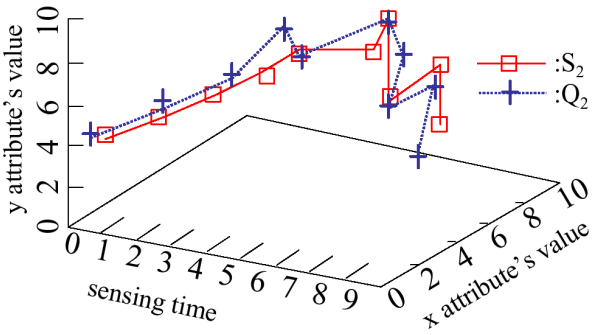
$$S=\begin{bmatrix}\text{측정시간}\\ \text{첫번째 속성}\\ \text{두번째 속성}\\ \cdot\\ \cdot\\ m\text{번째 속성}\end{bmatrix}\begin{bmatrix}t_1\\ S_1[1]\\ S_2[1]\\ \cdot\\ \cdot\\ S_m[1]\end{bmatrix}\begin{bmatrix}t_2\\ S_1[2]\\ S_2[2]\\ \cdot\\ \cdot\\ S_m[2]\end{bmatrix}\begin{bmatrix}t_3\\ S_1[3]\\ S_2[3]\\ \cdot\\ \cdot\\ S_m[3]\end{bmatrix}\cdots\begin{bmatrix}t_n\\ S_1[n]\\ S_2[n]\\ \cdot\\ \cdot\\ S_m[n]\end{bmatrix}$$

(그림 2)는 두 개의 속성으로 구성된 이차원 시퀀스의 예를 보여주고 있으며, 두 시퀀스 S_2, Q_2가 각각 측정시간, x-속성, y-속성 공간 상에 표시되어 있다.

여러 개의 속성으로 구성된 다차원 시퀀스 데이타의 유사 검색에 대한 기존 연구[13, 14, 15]는 다중속성 시퀀스에서 개별 속성의 시퀀스를 고려하지 못하고 있으며, 다음과 같이 시퀀스 내의 요소들 간의 거리 누적을 두 다중 속성 시퀀스 간의 유사도를 기반하고 있다.

$$D(S,Q)=\sum_{i=1}^{n}(\sum_{j=1}^{m}|S_m[i]-Q_m[i]|^2)^{1/2}$$

다중 속성 시퀀스 내의 요소들 간의 거리 누적에 기반한



(그림 2) 두 개의 속성을 가진 시퀀스의 예

유사도는 개별 속성 시퀀스에 대한 정보를 상실하게 되며, 궤도(trajectory) 정보와 같이 개별 시점에서 시퀀스 요소 간의 거리에 의미가 주어질 때 적합하다. 그에 비해 온도 및 습도로 구성된 기후 데이타와 같이 개별 속성의 시퀀스가 중요한 의미를 지니는 경우에는 적합하지 않다. 예를 들어 특정 지역과 온도의 변화가 ε_1 이내이고, 습도의 변화가 ε_2 이내인 유사한 지역을 검색할 때, 위의 정의된 유사도 함수는 개별 속성의 변화를 고려하고 있지 않기에 질의 처리가 어려운 단점이 있다.

본 논문에서는 기존의 다중속성 시계열 데이타베이스에 대한 연구에서 사용한 유사도 함수 대신 개별 속성 시퀀스 간의 거리에 기반한 다중 속성 시퀀스간의 유사도를 정의하고 그에 기반한 유사 검색 기법을 제안한 기존의 연구[5]를 확장하였다. 제안된 기법은 저차원 변환(lower dimensional transform)을 이용하여 다중속성 시계열을 하나의 속성을 가진 일차원 시계열로 사상한 후, 특징을 추출하게 된다. 제안된 기법은 검색 공간을 효율적으로 줄일 수 있으며, 착오누락 없이 유사한 시퀀스를 검색할 수 있음을 보장한다.

본 논문의 구성은 다음과 같다. 2절에서 시퀀스 검색과 관련된 연구에 대해 살펴보고, 3절에서 제안된 기법에 대해 설명한다. 4절에서 제안된 기법의 성능을 평가하며, 5절에서 결론을 맺는다.

2. 관련 연구

시계열 데이타의 빠른 검색과 매칭을 위해 다양한 기법들이 제안되어 왔다. 일반적인 방법은 고차원의 시퀀스를 차원 감소 기법을 사용하여 저차원으로 사상한 후 기존의 다차원 색인 구조를 이용하여 색인하는 방식이다. 시퀀스에서 특징을 추출하는 차원 감소 기법으로는 DFT(Discrete Fourier Transform)[1, 6], DWT(Discrete Wavelet Transform)[7], SVD (Singular Value Decomposition)[8], PAA(Piecewise Aggregate Approximation)[4, 9] 등이 있다.

시계열 데이타베이스에서 유사한 시퀀스를 검색하는 문제는 [1]에서 처음 제기되었다. F-index라 하는 색인 기법이 길이가 같은 시퀀스 데이타베이스에서 유사 검색 질의를 처리하기 위해 제안되었다. 이 기법은 DFT를 사용하여 차원 감소시킨 시퀀스를 R-tree[3]를 사용하여 색인하였으며, 유

사도로 유클리드 거리를 사용하였다.

차원 감소 기법으로 DFT 대신 DWT를 사용하는 연구가 [7]에서 수행되었다. DWT는 DFT에 비해 연산이 간단하며 효과적인 특징 추출이 가능하다. 그러나 DWT는 길이가 2의 지수인 시퀀스에 대해서 최적의 성능을 나타내는 단점이 있다.

SVD를 차원 감소 기법으로 이용한 연구는 [8]에서 수행되었다. SVD는 DFT나 DWT와 달리 전체 데이타 분포를 고려한 차원 감소 기법이기 때문에 효과적인 특징 추출이 가능하다. 그러나 계산 복잡도가 너무 높기 때문에 시간 공간적인 비용이 매우 큰 단점이 있다. 또한 새로운 시퀀스가 데이타베이스에 추가되는 경우 색인을 재구성을 해야 되는 단점을 가지고 있다.

시퀀스를 길이가 같은 여러 세그먼트로 나눈 후 각 세그먼트의 평균값을 특징으로 하는 차원 감소 기법이 [4, 9]에서는 제안되었다. 제안된 기법은 임의의 Lp 거리에 대해 적용할 수 있는 장점을 가지고 있다.

시퀀스간의 유사도로 유클리드 거리 대신 타임와핑(time warping)[10, 11, 12, 13] 거리를 사용한 연구도 다양하게 수행되어 왔다. 타임와핑은 길이가 다른 시퀀스를 비교하는데 있어 적합한 유사도의 척도를 제공한다. 그러나 타임와핑은 계산 복잡도가 높고, 삼각 부등식이 성립하지 않는 거리 함수이기 때문에 기존의 다차원 색인 구조로 색인하는 것이 쉽지 않다.

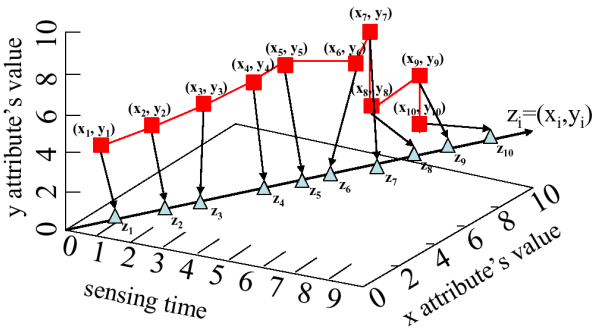
타임 와핑을 시퀀스 간의 유사도로 이용한 최신의 논문으로 [18, 19]가 있으며, 타임와핑의 하한 함수를 정의하고 이를 통해 기존의 공간 접근 기법을 통해 색인할 수 있음을 보이고 있다.

여러 개의 속성으로 구성된 다차원 시퀀스에 대한 연구는 [14, 15, 16]에서 수행되었으나, 다중속성 시퀀스에서 개별 속성의 시퀀스를 고려하지 않고 있다. 기존 연구에서 다중 속성 시퀀스 간의 유사도는 본질적으로 시퀀스 내의 요소들 간의 거리의 누적을 이용하고 있다. 따라서 다중 속성 시퀀스를 구성하는 개별 속성 시퀀스에 대한 정보를 상실하게 되며, 궤도(trajectory) 정보와 같이 개별 시점에서 시퀀스 요소 간의 거리에 의미가 주어질 때만 사용할 수 있으며, 기후 및 음악 데이타와 같이 개별 속성의 시퀀스가 중요한 의미를 지니는 경우에는 부적절하다.

특징 추출과 더불어 시퀀스 내의 서브시퀀스(subsequence) [20, 21] 검색을 위한 다양한 연구가 수행되고 있다.

3. 제안된 기법

본 논문에서는 다중속성 시계열 데이타베이스에서 질의 시퀀스와 유사한 시퀀스를 빠르게 찾는 문제에 초점을 두고 있다. 우선 다차원 시퀀스와 유사도를 정의하고, 그 이후에 효율적인 유사 검색 기법을 제안하며, 착오 누락이 없음을 보인다. 다중속성 시퀀스는 여러 개의 속성 시퀀스로 구성되어 있으며, 속성 시퀀스 측면에서 유사도를 고려하는 것이 적절할 수 있다. 본 논문에서는 두 개의 다중속성 시퀀스 S, Q 사이의 유사도를 다음과 같이 속성 시퀀스 간의 유클리드 거리의 가중 합을 이용한 통합 유사도 모델을 사용하였다.



(그림 3) 저차원 변환의 예

정의 2 (통합 유사도 모델) m개의 속성을 가지며 길이가 n인 두 시퀀스 S, Q간의 유사도는 다음과 같이 속성 시퀀스 간의 유클리드 거리의 가중 합으로 정의된다.

$$D_{total}(S,Q)=\sum_{j=1}^{m}\lambda_j\left(\sum_{i=1}^{n}|S_m[i]-Q_m[i]|^2\right)^{1/2}$$

λ_j (j = 1, 2, ..., m)는 각 속성에 대한 가중치를 나타낸다. 이러한 가중치는 각 속성의 절대치가 서로 다른 경우 정규화하기 위해서 또는 사용자의 개별 속성 시퀀스에 대한 선호도를 지원하기 위해 사용된다. 두 다중속성 시퀀스 간의 통합 유사도가 주어진 허용 한계 ε 이내이면 두 시퀀스는 서로 유사하다고 볼 수 있으며, 이러한 통합 유사도에 기반하여 유사 검색 질의를 처리 할 수 있게 된다. 통합 유사도는 개별 속성시퀀스에 대한 유사 검색 질의조건이 주어질 때에도 사용 가능하다. 즉 개별 속성에 대한 유사 범위를 가중 합하여 질의 조건으로 주어서 처리하게 된다.

본 절에서는 저차원 변환을 통해 다중속성 시퀀스를 하나의 속성으로 된 일차원 시퀀스로 변환하여, 변환된 시퀀스에서 특징하는 것에 대해 기술한다. 여기에서 중요한 점은 저차원 변환에 의해 생성된 시퀀스 간의 거리가 다중차원 시퀀스 간의 거리보다 작거나 같아야 한다는 것이다. 즉 $D_{lower}(L(S),L(Q)) \le D_{total}(S,Q)$를 만족한다면 착오누락이 없음을 보장하게 된다. 본 논문에서는 저차원 변환으로 측정 시간에서의 각 속성 요소 값의 합을 사용하였다[1]. (그림 3)은 저차원 변환의 예를 보여주고 있다.

측정 시간에서의 각 속성 요소 값의 합을 사용한 저차원 변환이 착오 누락이 없음은 다음과 같이 보일 수 있다. 저차원으로 변환된 두 시퀀스 L(S), L(Q) 사이의 거리는 다음과 같이 정의된다.

$$D_{lower}(L(S),L(Q))=\left(\sum_{i=1}^{n}|(\lambda_1 S_1[i]+\cdots+\lambda_m S_m[i])-(\lambda_1 Q_1[i]+\cdots+\lambda_m Q_m[i])|^2\right)^{1/2}$$

1) 하한 조건(lower bound condition)을 만족한다면 다양한 저차원 변환이 사용될 수 있다.

착오 누락이 없음을 보이기 위해서는 저차원 변환된 시퀀스 간의 거리가 원래의 다중속성 시퀀스 간의 거리의 하한임을 보여야 한다.

정리 1. 저차원 변환된 시퀀스 간의 거리 $D_{lower}(L(S), L(Q))$는 통합 유사도 $D_{total}(S,Q)$의 하한을 만족하며, 착오 누락이 없음을 보장한다.

증명. $D_{lower}(L(S), L(Q)) \le D_{total}(S,Q)$를 만족함을 삼각 부등식을 이용하여 증명한다.

$\overrightarrow{V_j}$를 j번째 속성의 시퀀스를 나타내는 벡터라고 한다면 다음과 같이 둘 수 있다. 여기서 단위벡터 $\overrightarrow{e_i}$, (i=1, 2, ..., n)는 i-차원에 대한 기본 벡터를 나타낸다.

$$\overrightarrow{V_j} = \lambda_j(\sum_{i=1}^{n}(s_j[i]-Q_j[i])\overrightarrow{e_i})$$

삼각 부등식을 이용하여, 다음과 같이 정리 1은 증명된다.

$$\begin{aligned}|\overrightarrow{V_1}+\cdots+\overrightarrow{V_m}|^2 &= |\overrightarrow{V_1}|^2+\cdots+|\overrightarrow{V_m}|^2+2|\overrightarrow{V_1}||\overrightarrow{V_2}|\cos\theta_1 \\ &\quad +\cdots+2|\overrightarrow{V_{m-1}}||\overrightarrow{V_m}|\cos\theta_{m(m-1)/2-1} \\ &\le |\overrightarrow{V_1}|^2+\cdots+|\overrightarrow{V_m}|^2+2|\overrightarrow{V_1}||\overrightarrow{V_2}| \\ &\quad +\cdots+2|\overrightarrow{V_{m-1}}||\overrightarrow{V_m}| \\ &= (|\overrightarrow{V_1}|+|\overrightarrow{V_2}|+\cdots+|\overrightarrow{V_m}|)^2\end{aligned}$$ □

위의 정리에 의해 질의로 주어진 다중속성 시퀀스와 유사도가 ε이내인 시퀀스를 찾기 위해 저차원 변환을 수행하여 하나의 속성을 가진 일차원 시퀀스를 생성하고, 생성된 일차원 시퀀스에서 DWT 등의 특징 추출 기법을 사용하여 저차원 특징을 추출한다. 추출된 특징을 이용하여 R-tree와 같은 다차원 색인 구조를 이용하여 시퀀스를 색인하여 검색에 이용한다. 알고리즘 1은 제안된 검색 기법을 기술하고 있다.

질의 처리 과정은 공간접근 기법을 통한 후보 검색과 후처리과정을 통한 착오 해답 과정의 두 단계로 구성되어 있다. 후처리 과정에서 실제 통합 유사도가 주어진 허용 한계를 벗어나는 착오 해답(false alarms)을 제거하고, 허용 한계 이내인 유사한 시퀀스들만을 사용자에게 질의 결과로 반환한다.

알고리즘 1. 저차원 변환 질의 처리

```
Input : query sequence Q, threshold ε
Output: data sequences within threshold ε
Begin
result ← NULL; candidate ← NULL;

// 저차원 변환후 특징을 추출하여 특징 공간으로 사상
FQ ← FeatureExtraction(LowerDimensionTransform(Q));

// 색인을 이용하여 후보 시퀀스 검색
candidate ← candidate U IndexSearching(FQ, ε);

// 착오 해답(false alarms) 제거
for(i=1; i < size of candidate; i++)
        if(ComputeActualDistance(Ci, Q) ≤ ε ) // Ci ∈ candidate
                result ← Ci U result ;
        else reject Ci ;
return result ;
End
```

4. 성능 평가

이 절에서는 제안된 기법의 성능을 분석하기 위한 실험 결과를 제시한다. 제안된 기법의 성능을 평가하기 위해 순차 검색과 검색 공간 비율(search space ratio) 측면에서 비교해 보았다. 개별속성 시퀀스 간의 거리를 이용하여 다중속성 시퀀스 간의 유사도를 정의하는 기존의 연구가 없기에 여기에서는 순차 검색과 성능을 비교하였다. 검색 공간 비율은 다음과 같이 정의된다.

$$\text{검색 공간 비율} = \frac{\text{후보 시퀀스의 개수}}{\text{전체 시퀀스의 개수}}$$

검색 공간 비율은 특징벡터의 성능과 관련이 있으며, 색인 구조, 하드웨어, 소프트웨어 등과 같은 구현 요소와는 무관하며 이러한 성능 평가 방법은 [9, 17] 등에 나타나 있다. 먼저 실험 환경에 대해 설명하고, 그 다음에 실험 결과를 제시하였다.

4.1 실험 환경

실험을 위해 2중 속성 시퀀스 및 3중 속성 시퀀스를 다음과 같은 random walk model[1]에 따라 생성하였다. 각 속성 시퀀스의 초기값 $S_m[0]$는 균일분포(uniform distribution)를 따르며, -500과 500 사이의 값을 생성하며, 이전 값과의 변이차 z_i는 균일분포(uniform distribution)을 따르며 정수 값을 발생시킨다.

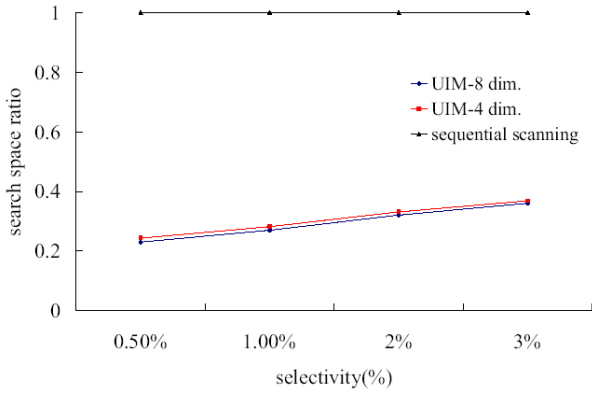
$$S_m[i] = S_m[i-1] + \alpha \cdot z_i$$

여기에서 $S_m[0] \sim U(-500, 500)$, $z_i = N(-500, 500)$, $\alpha = 0.05$

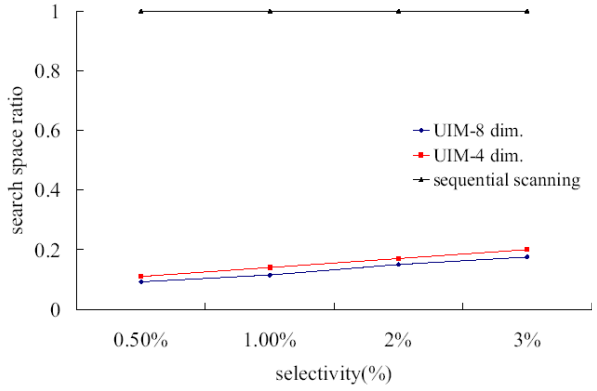
가중치 요소 λ_j는 사용자의 편의에 따라 각 속성 시퀀스의 값의 범위 및 특성을 고려하여 조정될 수 있으며, 본 실험에서는 편의를 위해 1로 설정하였다. 또한 일차원으로 변환된 시퀀스에서 특징을 추출하기 위해 DWT를 사용하였다. DWT 외에 DFT나 PAA와 같은 특징 추출 기법도 사용가능하다. 여기에서 중요한 점은 특징의 차원이며, 본 실험에서는 4차원, 8차원의 특징벡터를 이용하였다. 4차원 및 8차원의 특징을 추출한 것은 R-tree와 같은 기존의 공간 접근 기법이 효율적으로 작동하는 차원으로 알려져 있기 때문이다. 질의 시퀀스도 데이타 시퀀스와 동일한 방식으로 생성하였으며, 시퀀스 데이타베이스에 대해 50번씩 질의를 수행하여, 평균적인 검색 공간 비율을 측정하였다.

4.2 실험 결과

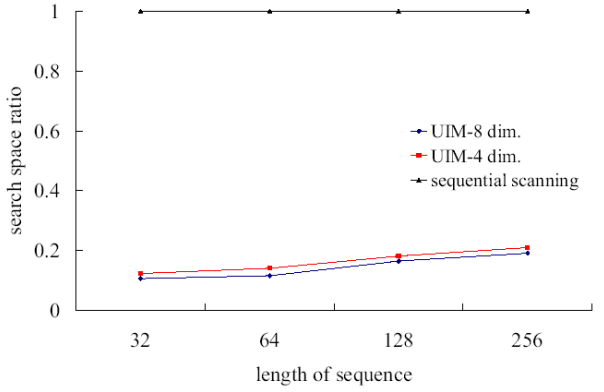
우선 허용 한계를 변화시키면서 실제 통합 유사도 계산이 필요한 후보 시퀀스의 개수에 대해 제안 기법과 순차 검색 기법을 비교해 보았다. 이 실험에서는 길이가 64인 2중 속성 시퀀스 및 3중 속성 시퀀스를 각각 10,000개씩 생성하였다. 허용 한계는 각 데이타셋에 대하여 평균 결과 선택도가 0.5%, 1%, 2%, 3%가 되도록 설정하였다. (그림 4)와 (그림 5)는 실



(그림 4) 허용 한계 변화에 따른 3중 속성 시퀀스에서 검색 공간 비율



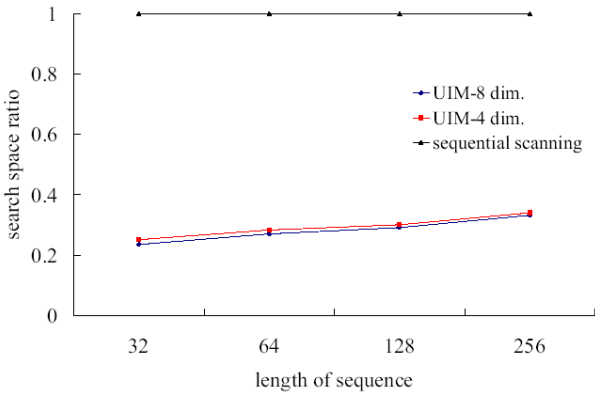
(그림 5) 허용 한계 변화에 따른 2중 속성 시퀀스에서 검색 공간 비율



(그림 6) 길이 변화에 따른 2중 속성 시퀀스에서 검색 공간 비율

험 결과를 보이고 있으며, 제안된 기법에 의해 검색 공간이 효율적으로 줄어들고 있음을 알 수 있다. 실험에서 제안된 기법은 UIM(unified index method)란 이름과 특징 차원의 정보를 덧붙여 정의한다.

제안 기법의 확장성을 평가하기 위해 시퀀스의 길이를 변경하면서 검색 공간 비율을 측정해 보았다. 본 실험에서는



(그림 7) 길이 변화에 따른 3중 속성 시퀀스에서 검색 공간 비율

결과 선택도가 1%가 되도록 허용 한계를 설정하였다. 시퀀스의 길이를 32에서 256까지 증가시키면서 검색 공간 비율을 측정하였으며, 2중 속성 및 3중 속성 시퀀스를 10,000개씩 각각 생성하였다. (그림 6)과 (그림 7)은 실험 결과를 보이고 있으며, 제안된 기법에 의해 시퀀스의 길이가 길어지더라도 검색 공간이 효율적으로 줄어들고 있음을 알 수 있다.

5. 결 론

본 논문에서는 다중속성 시퀀스 간의 통합 유사도를 기반으로 유사 검색 기법에 대해 기술하였다. 기존의 연구는 다중속성 시퀀스를 구성하는 개별 속성의 시퀀스를 고려하지 못하였으나, 본 논문에서는 개별 속성 시퀀스 측면에서 다중속성 시퀀스간의 유사도 모델을 정의하였으며, 이를 기반으로 다중속성 시계열 데이타베이스에서의 유사 검색 기법을 제안하였다. 제안된 기법은 저차원 변환 기법을 통해 다중속성 시퀀스를 일차원 시퀀스로 사상한 후, 특징을 추출하여 색인을 구성하여 효율적인 유사검색을 수행 할 수 있다. 또한 제안된 기법이 착오누락 없이 정확하게 질의를 처리할 수 있음을 증명하였으며, 실험을 통해 효율적으로 검색 공간을 줄이며, 시퀀스 길이 변화에 대해서도 확장성이 있음을 보였다.

참 고 문 헌

[1] Rakesh Agrawal, Christos Faloutsos and Arun N. Swami, “Efficient Similarity Search in Sequence Databases,” Proceedings of the International Conference of Foundations of Data Organization and Algorithms, pp.69-84, 1993.

[2] Antonin Guttman, “R-trees: A Dynamic Index Structure for Spatial Searching,” Proceedings of ACM SIGMOD International Conference on Management of Data, pp.47-57, 1984.

[3] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider and Bernhard Seeger, “The R*-tree: An Efficient and Robust

Access Method for Points and Rectangles," Proceedings of ACM SIGMOD International Conference on Management of Data, pp.322-331, 1990.

[4] Byoung-Kee Yi, Christos Faloutsos, "Fast Time Sequence Indexing for Arbitrary Lp Norms," Proceedings of International Conference on Very Large Data Bases, pp. 385-394, 2000.

[5] Sangjun Lee, Bumsoo Kim, Sukho Lee, "Efficient Range Search Method for Multi-dimensional Sequence Databases," KISS Journal, Vol. 26(5), pp.613-620, 1999.

[6] Davood Rafiei, Alberto O. Mendelzon, "Similarity-Based Queries for Time Series Data," Proceedings of ACM SIGMOD International Conference on Management of Datae, pp.13-25, 1997.

[7] Kin-pong Chan, Ada Wai-chee Fu, "Efficient Time Series Matching by Wavelets," Proceedings of International Conference on Data Engineering, pp.126-133, 1999.

[8] Flip Korn, H. V. Jagadish, Christos Faloutsos, "Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences," Proceedings of ACM SIGMOD International Conference on Management of Data, pp.289-300, 1997.

[9] Eamonn J. Keogh, Kaushik Chakrabarti, Sharad Mehrotra, Michael J. Pazzani, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases," Proceedings of ACM SIGMOD International Conference on Management of Data, pp.151-162, 2001.

[10] Byoung-Kee Yi, H. V. Jagadish, Christos Faloutsos, "Efficient Retrieval of Similar Time Sequences Under Time Warping," Proceedings of International Conference on Data Engineering, pp.201-208, 1998.

[11] Sangwook Kim, Sanghyun Park and W. Chu, "An Index-based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," Proceedings of International Conference on Data Engineering, pp.607-614, 2001.

[12] Eamonn J. Keogh, "Exact Indexing of Dynamic Time Warping," Proceedings of International Conference on Very Large Data Bases, pp.406-417, 2002.

[13] Sanghyun Park, Wesley W. Chu, Jeehee Yoon, Chihcheng Hsu, "Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases," Proceedings of International Conference on Data Engineering, pp.23-32, 2000.

[14] Seok-Lyong Lee, Seok-Ju Chun, Deok-Hwan Kim, Ju-Hong Lee and Chin-Wan Chung, "Similarity Search for Multidimensional Data Sequences," Proceedings of International Conference on Data Engineering, pp.599-608, 2000.

[15] Michail Vlachos, G.Kollios and Dimitrios Gunopulos, "Discovering Similar Multidimensional Trajectories," Proceedings of International Conference on Data Engineering, pp.673-684, 2002.

[16] Tamer Kahveci, Ambuj Singh and Aliekber Gurel, "Similarity Searching for Multi-attribute Sequences," Proceedings of International Conference on Scientific and Statistical Database Management, pp.175-184, 2002.

[17] Joseph M. Hellerstein, Elias Koutsoupias, and Christos H. Papadimitriou, "On the Analysis of Indexing Schemes," Proceedings of ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp.249-256, 1997.

[18] Ada Wai-Chee Fu, Eamonn J. Keogh, Leo Yung Hang Lau, Chotirat (Ann) Ratanamahatana, "Scaling and Time Warping in Time Series Querying," Proceedings of International Conference on Very Large Data Bases, pp.649-660, 2005.

[19] Eamonn J. Keogh, Li Wei, Xiaopeng Xi, Sang-Hee Lee, Michail Vlachos, "LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures," Proceedings of International Conference on Very Large Data Bases, pp.882-893, 2006.

[20] Sang-Wook Kim, Dae-Hyun Park, Heon-Gil Lee, "Efficient Processing of Subsequence Matching with the Euclidean Metric in Time-series Databases," Information Process. Letters, Vol. 90(5), pp.253-260, 2004.

[21] Yang-Sae Moon and Jinho Kim, "A Single Index Approach for Time-Series Subsequence Matching that Supports Moving Average Transform of Arbitrary Order," Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.739-749, 2006.

이 상 준

e-mail : sangjun@ssu.ac.kr

1996년 서울대학교 컴퓨터공학과(학사)

1998년 서울대학교 대학원 컴퓨터공학과 (공학석사)

2004년 서울대학교 대학원 전기컴퓨터공학부 (공학박사)

2004년~2005년 자동제어특화연구센터 연구원

2005년~현 재 숭실대학교 컴퓨터학부 조교수

관심분야 : 멀티미디어, 데이타베이스, 데이타마이닝, P2P 시스템 등