# 시계열 거리 계산에서 미리 버림 효과의 최대화*

이정곤°, 김상필, 문양세
강원대학교 컴퓨터과학과
e-mail : jglee, spkim, ysmoon@kangwon.ac.kr

# Maximizing the Early Abandon Effect in Time-Series Distance Computation

Jeong-Gon Lee, Sang-Pil Kim, Yang-Sae Moon
Dept. of Computer Science, Kangwon National University

**요 약**

본 논문에서는 유사 시퀀스 매칭에서 미리 버림 계산의 효율적인 방법을 제안한다. 미리 버림은 유사 시퀀스 매칭에서 유클리디안 거리 계산 도중 거리 계산 값이 허용치보다 큰 경우 나머지 거리 계산을 하지 않는 방법이다. 기존의 방법은 시퀀스 첫 엔트리를 시작으로 하여 유클리디안 거리 계산을 진행한다. 이 방법은 데이터 고려 없이 계산이 진행되기 때문에 데이터의 특성에 따라 효과가 크게 다른 점을 보인다. 본 논문에서는 미리 버림의 효과를 최대화 시키기 위해 유클리디안 거리 계산 시작점을 오프셋이라 정의하고, 이를 데이터 특성에 맞게 조절하는 방법을 제안한다. 실험 결과, 제안한 오프셋 조절 미리 버림 방법이 대용량의 데이터 베이스 기반 시스템에서 기존 기법에 비해 좋은 성능 향상시킨 것으로 나타났다.

## 1. 서론

최근 대용량 시계열 데이터베이스 대상의 시계열 매칭(time-series matching)에 관한 연구가 활발하게 이루어져 왔다[1, 2, 3]. 또한, 최근에는 여러 응용에서 시계열 매칭 연구가 활용되고 있다. 본 논문에서는 시계열(시퀀스)매칭 거리 계산의 성능 향상에 대해서 다루고자 한다.

유클리디안 거리[4] 계산 방법은 시퀀스 매칭에서 많이 사용하는 거리 계산 방법 중 하나이다. 두 시퀀스의 거리를 측정하는 유클리디안 거리 계산 방법은 다음과 같이 정의한다.

**정의 1:** 길이가 $n$ 인 두 시퀀스 $Q=q_0,q_1,...,q_{n-1}$ 과 $S=s_0,s_1,...,s_{n-1}$ 두 시계열의 유클리디안 거리 $D(Q,S)$ 는 다음 식(1)과 같이 정의 한다.

$$D(Q,S)=\sqrt{\sum_{i=0}^{n-1}(q_i-s_i)^2} \qquad (1)$$

두 시퀀스 0 번째부터 $n$-1 까지의 거리 차의 합을 구하는 것이 유클리디안 거리이며, 거리 계산의 시작 위치를 우리는 오프셋(offset)이라고 한다. □

유클리디안 거리 계산 과정에서 중간까지의 거리 값이 주어진 허용치보다 커지면 계산을 중단하는 방법을 미리 버림(early abandon)[5]이라 한다. 즉, 거리 계산이 끝나지 않은 경우에도 유사하지 않다고 판단하는 경우이다. 따라서, 불필요한 계산을 하지 않기 때문에 성능의 향상을 가져 올 수 있다. 그림 1 은 미리 버림 효과에 대한 예를 보여준다. 본 논문에서는 유클리디안 거리 계산에서 오프셋 조정을 통한 성능 향상에 논의한다.

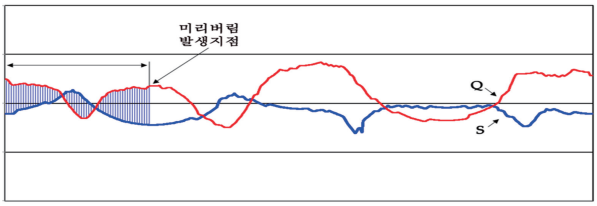


그림 1. 미리 버림을 이용한 유클리디안 거리 계산 과정.

## 2. 관련 연구 및 기존 알고리즘

시계열 데이터는 일정한 간격으로 수집된 데이터 시퀀스이다. 이러한 시계열 데이터의 예로는 주식, 날씨 데이터, 환율 변동 데이터 등이 있으며, 그 외 이미지, 음성, 동영상 등의 다양한 분야에서도 응용되고 있다. 이러한 시계열 데이터의 처리에는 저차원 변환[6], 프라이버시 보호[5], 유사 시퀀스 매칭[1] 등의 다양한 연구가 활발하게 진행되고 있다. 본 논문에서는 유사 시퀀스 매칭 기법에 대해 다루며, 유사 시퀀스 매칭은 사용자에 의해 주어진 질의 시퀀스와 시계열 데이터 베이스에 저장된 데이터 시퀀스를 비교하여, 질의 시퀀스와 유사한 데이터 시퀀스를 찾는 작업이다.

유사 시퀀스 매칭에는 두 시퀀스의 거리를 측정하는 방법이 자주 사용된다. 이러한 유사 시퀀스 매칭 거리 측정 방법에는 유클리디안 거리, DTW[3] 등의 거리 측정 방법이 있으며, 본 논문에서는 유클리디안 거리 계산에 초점을 맞춘다. 유클리디안 거리는 두

* 이 논문은 2010 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2010-0002518)

시퀀스 $Q$, $S$ 사이의 거리가 사용자가 제시한 허용치인 ε이하이면 두 시퀀스가 유사하다고 정의한다. 이러한 유클리디안 거리 계산 과정에서 중간까지의 거리 값이 주어진 허용치보다 커지면 계산을 중단하는 방법을 미리 버림이라고 한다. 이러한 기법은 시계열 데이터 처리에서 유사성을 비교하는 모든 처리 분야에 적용할 수 있다.

미리 버림을 사용한 유사 시퀀스 매칭의 기본 알고리즘 *BasicOffset*()는 그림 2 와 같다. 그림에서 보듯이, 데이터 시퀀스 $S$ 와 질의 시퀀스 Q 의 유클리디안 거리를 계산하며, 계산 중간 거리 값이 허용치보다 커지면 유사하지 않은 것으로 판단(라인 4)한다. 기존 미리 버림 방법은 시퀀스 첫 오프셋에서 시작해 유클리디안을 계산하는 방법으로 본 논문에서는 오프셋 위치를 변경해 미리 버림의 효과를 향상 시키는 방법을 제시한다.

```
Function BasicOffset(query sequence Q, data sequence S, tolerance ε)
1.  sqdist := 0;
2.  for i:=0 to (n–1) do
3.      sqdist := sqdist + (q_i – s_i)^2 ;
4.      if (sqdist > ε²) en return √sqdist ;
5.  end-for
6.  return √sqdist ;
```

그림 2. 미리 버림을 수행하는 기존 알고리즘.

## 3. 최대값 오프셋 선택

본 절에서는 비교 대상인 두 시퀀스의 거리 차이를 고려한 최대값 오프셋 선택 방법을 제안한다. 이는 유사 시퀀스 매칭에서 질의 시퀀스와 데이터 시퀀스 간의 거리 차이가 큰 오프셋부터 거리를 계산하면 미리 버림 효과가 커질 것이라는 점에 착안한 것이다. 직관적으로 거리 비교하는 두 시퀀스의 거리 차이는 시퀀스 내 최대값 혹은 최소값에서 발생할 가능성이 높다. 따라서 본 절에서는 시퀀스 내 최대값 오프셋을 선택하여 유클리디안 거리를 계산한다. 그림 3 이 최대값 오프셋 기반 미리 버림 사용 유클리디안 거리 계산을 나타낸 것이다.

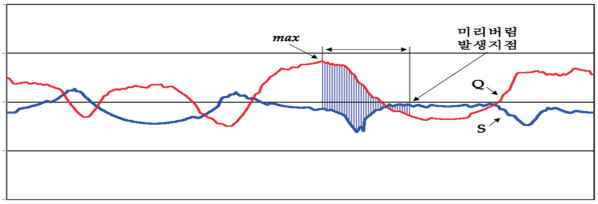


그림 3. 질의 시퀀스의 최대값 거리 계산 과정.

그림 4 는 최대값 오프셋 기반 미리 버림 사용 유클리디안 거리 계산 알고리즘이다. 그림 4 에서 라인 2 는 최대값을 갖는 오프셋을 선택하는 과정이고, 오프셋을 선택한 이후의 나머지 과정(라인 3-7)은 유클리디안 거리를 계산하는 과정이다. 알고리즘에서 최대값 대신에 최소값을 사용할 수도 있으나, 그 효과는 동일(혹은 유사)하므로 본 논문에서는 최대값의 경우만 고려한다.

```
Function QueryMax(query sequence Q, data sequence S, tolerance ε)
1.  sqdist := 0;
2.  j := an offset where q_j is the maximum entry in Q;
3.  for i:=0 to (n–1) do
4.      sqdist := sqdist + (q_((j+i)%n) – s_((j+i)%n))^2 ;
5.      if (sqdist > ε²) en return √sqdist ;
6.  end-for
7.  return √sqdist
```

그림 4. 질의 시퀀스의 최대값 오프셋을 선택한 계산 알고리즘.

알고리즘 *QueryMax*()에서 최대값을 찾는 과정(라인 2)이 오버헤드로 작용할 수 있다. 그러나, 이는 매 거리 계산마다 이루어지는 것이 아니라, 질의 시퀀스 $Q$ 에 대해서 한번만 수행되는 과정으로서, 비교 대상인 데이터 시퀀스가 많은 일반적인 환경에서는 충분히 무시할 수 있는 연산으로 볼 수 있다

## 4. 최대/최소값 오프셋 선택

본 절에서는 최대/최소값 오프셋 기반 미리 버림을 사용한 유클리디안 거리 계산 방법을 제안한다. 이는 데이터 특성에 따라 최대값 오프셋 선택이 미리 버림 효과를 크게 발휘할 수도 있고, 최소값 오프셋 선택이 미리 버림 효과를 크게 발휘할 수도 있기 때문이다. 그림 5 는 최대/최소값 오프셋 기반 미리 버림 사용 유클리디안 거리 계산을 나타낸 것이다.

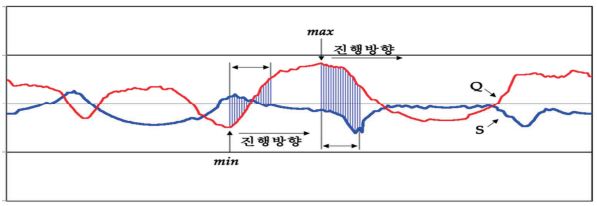


그림 5. 질의 시퀀스의 최대값과 최소값의 거리 계산 과정.

그림 6 은 최대/최소값 오프셋 선택을 사용한 유클리디안 거리 계산 알고리즘인 *QueryMinMax*()을 나타낸다. 그림에서 라인 2 와 3 은 각각 최소값과 최대값을 갖는 엔트리의 오프셋을 찾는 과정으로, 질의 시퀀스에 대해서 단 한번만 수행된다. 라인 5-8 은 최소값 오프셋(*min*)에서 시작하여 거리를 계산하는 과정이다.

```
Function QueryMinMax(query sequence Q, data sequence S, tolerance ε)
1.  sqdist := 0;
2.  min := an offset where q_min is the minimum entry in Q;
3.  max := an offset where q_max is the maximum entry in Q;
4.  for i:=0 to (n–1) do
5.      if ((min+i)%n < max) then begin
6.          sqdist := sqdist + (q_((min+i)%n) – s_((min+i)%n))^2 ;
7.          if (sqdist > ε²) then return √sqdist ;
8.      end-if
9.      if ((max+i)%n < min) then begin
10.         sqdist := sqdist + (q_((max+i)%n) – s_((max+i)%n))^2 ;
11.         if (sqdist > ε²) then return √sqdist ;
12.     end-if
13. end-for
14. return √sqdist ;
```

그림 6. 질의 시퀀스의 최소값 및 최대값 오프셋을 사용하는 계산 알고리즘.

## 5. 성능평가

### 5.1 실험 환경 및 데이터

실험에서는 총 세가지 데이터 집합을 사용하였다. 첫 번째 데이터 집합은 전자회로 접압 측정 데이터로, 길이 100 인 시계열 데이터 1000 개로 구성되어 있다. 데이터 집합의 이름을 *LOG_DATA* 라 한다. 두 번째 데이터는 미항공우주국(NASA)에서 측정한 웨이블릿 데이터로, 길이 20000 인 데이터 27 개로 구성되어 있다. 이 데이터 집합을 *WAVE_DATA* 라 한다. 마지막으로 교차로 차량 통행량을 측정한 데이터 집합으로 길이가 1000 인 시계열 데이터 154 개로 구성되어 있다. 이 데이터 집합을 *CROSS_DATA* 라 한다.

실험은 기존 방법인 오프셋 0 유클리디안 거리 계산 방법인 basic 과 본 논문에서 제안한 최대값 오프셋 유클리디안 거리 계산 방법인 max, 그리고 최대/최소값 오프셋 유클리디안 거리 계산 방법인 minmax 세가지 알고리즘을 대상으로 하였다. 실험은 10 개 질의 시퀀스에 대한 유클리디안 거리 계산 횟수의 평균을 측정값으로 사용하였다. 실험에 허용치는 데이터 집합의 유클리디안 거리 평균을 측정한 후, 유클리디안 거리를 일정하게 나누어 사용하였다.

### 5.2 실험 결과

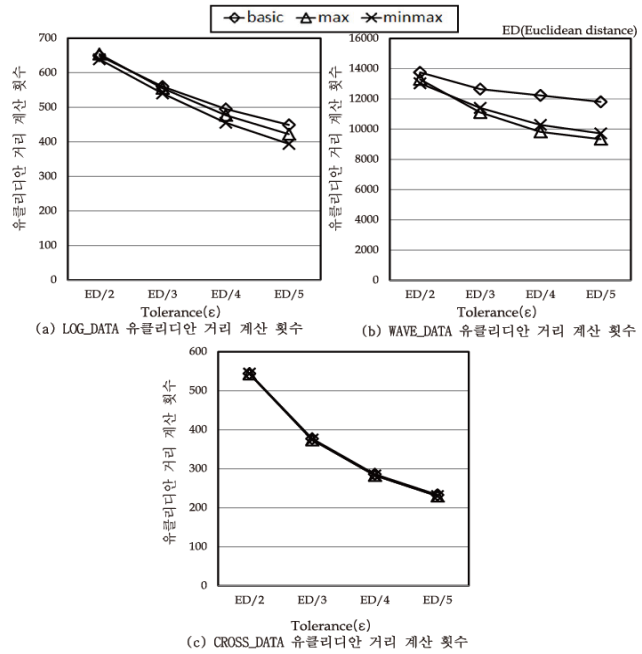


그림 7. 데이터 집합 별 유클리디안 거리 계산 횟수.

그림 7 는 허용치 변화에 따른 세가지 데이터 집합의 유클리디안 거리 계산 횟수를 측정한 그래프이다. 그래프가 전체적으로 오른쪽으로 가면서 값이 작아지는 것은 허용치가 작아짐에 따라 미리 버림이 효과가 나타나 유클리디안 거리 계산 횟수를 줄이기 때문이다. 우선 그림 7(a)를 보면, 제안한 방법들이 기존의 방법에 비해 유클리디안 거리 계산횟수를 전체적으로 줄인 것을 확인할 수 있다. 특히 minmax 방법이 기존 basic 방법에 비해 계산 횟수를 많이 줄인 것을 볼 수 있다. 이는 LOG_DATA 의 경우 데이터의 변화 폭이 넓게 형성되어 있기 때문이다. 그림 7(b)를 보면, 전체적인 그래프 형태는 그림 7(a)보다 제안한 방법이 확실하게 성능향상을 한 것을 알 수 있다. 하지만, minmax 방법 보다 max 방법이 효과가 더 좋게 나타난 것을 볼 수 있다. 이는 WAVE_DATA 의 경우 데이터 값이 고르게 분포 되어있고 간혹 높은 수치 값 분포가 있어 최소값 오프셋에서 거리 차이가 없어, 최대값 오프셋에서만 진행하는 거리 계산 값이 커 max 방법이 좋은 결과를 보인 것이다. 그리고 마지막으로 그림 7(c)의 경우, 기존 방법과 제안한 방법들이 모두 비슷한 결과를 보인다. 이것은 데이터 값 분포가 큰 변화가 없을 경우에 나타난다. 그 이유는 최대/최소값 오프셋 값과 평균 분포된 값과 차이가 크지 않아 기존 0 오프셋이나 최대/최소값 오프셋 거리 계산 값이 비슷하게 계산되기 때문이다. 그림 7 를 종합해 보면, 데이터의 특성에 따라 성능의 차이가 있는 것을 알 수 있으나, 전체적으로 제안한 방법이 좋은 성능향상을 보였다.

## 6. 결론

본 논문에서는 유사 시퀀스 매칭에 있어, 미리 버림을 사용한 유클리디안 거리 계산의 효과를 최대화하는 방법을 제안하였다. 이를 위해, 우선 유클리디안 거리 계산 시작점을 오프셋이라 정의하였다. 우선, 최대값 오프셋 기반 유클리디안 거리 알고리즘을 제안하였다. 그리고 최대/최소값 오프셋 기반 유클디안 거리계산 알고리즘을 제안하였다. 또한 세가지 데이터 집합에 대한 실험을 통해 기존의 미리 버림 방법보다 효과적인 것을 확인하였다.

## 참고문헌

[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] Keogh, E., “Exact Indexing of Dynamic Time Warping,” In Proc. *The 28th Int’l Conf. on Very Large Data Bases*, Hong Kong, pp. 406-417, Aug. 2002.

[3] Moon, Y.-S., Kim, H.-S., Kim, S.-P., and Bertino, E., “Publishing Time-Series Data under Preservation of Privacy and Distance Orders,” In proc. *21th Int’l Conf. on Database and Expert Systems Application*, Bilbao, Spain, pp. 17-31, Sept. 2010.

[4] Breu, H., Gil, J., Kirkpatrick, D., and Werman, M., “Linear time Euclidean distance transform algorithms,” IEEE Tans. on *Pattern Analysis and Machine Intelligence*, Vol. 17, Issue 5, pp. 529-533, May 1995

[5] Keogh, E. j., Wei, L., Xi, X., Vlachos, M., Lee, S.-H., and Protopapas, P., “Supporting Exact Indexing of Arbitrarily Rotated Shapes and Periodic Time Series under Euclidean and Warping Distance Measures,” *The VLDB Journal*, Vol. 18, No. 3, pp. 611-630, June 2009.

[6] Moon, Y.-S., and Kim, J.-H., "Hybrid Lower-Dimensional Transformation for Similar Sequence Matching," *IEICE Transactions on Information and Systems*, Vol. E92.D, Issue 3, pp. 541-544, March 2009.