http://dx.doi.org/10.5392/JKCA.2011.11.12..052

최소 DTW 거리 기반의 데이터 시퀀스 색인 기법

Sequence Data Indexing Method based on Minimum DTW Distance

길기정*, 송석일*, 송재종**, 이석필**, 장세진**, 이종설**
충주대학교 컴퓨터공학과*, 한국전자부품연구원**

Ki-jeong Khil(setoutroad@gmail.com)*, Seokil Song(sisong@cjnu.ac.kr)*,
Chai-Jong Song(jcsong@keti.re.kr)**, Seok-Pil Lee(lspbio@keti.re.kr)**,
Sei-Jin Jang(sjjang@keti.re.kr)**, Jong-Seol Lee(leejs@keti.re.kr)**

요약

이 논문에서는 시퀀스 데이터베이스에서 효과적인 유사 검색을 지원하기 위한 색인 기법을 제안한다. 제안하는 색인 기법에서는 데이터 시퀀스에 대한 필터링 효과를 얻기 위해, 최소 DTW 거리를 새롭게 제안한다. 최소 DTW 거리는 유사한 데이터 시퀀스 그룹과 질의 시퀀스 사이의 최소거리를 측정하는 방법이다. 최소 DTW 거리는 계층적인 색인 구조를 통해서 시퀀스 데이터베이스를 필터링하면서 유사도 검색을 수행할 수 있도록 한다. 마지막으로, 실험을 통해서 제안하는 방법의 우수성을 입증한다.

■ 중심어 : | 시계열 | DTW | 색인 | 시퀀스 |

Abstract

In this paper, we propose an indexing method to support efficient similarity search for sequence databases. We present a new distance measurement called minimum DTW distance to enhance the filtering effects. The minimum DTW distance is to measure the minimum distance between a sequence data and the group of similar sequences. It enables similarity search through hierarchical index structure by filtering sequence databases. Finally, we show the superiority of our method through some experiments.

■ keyword : | Time Series | DTW | Index | Sequence |

I. 서 론

데이터 시퀀스는 상업분야, 과학분야, 공학분야 등에서 중요하게 취급되고 있으며, 의사 결정 및 정책 수립에 중요한 역할을 수행한다. 데이터 시퀀스를 다루는 실세계 응용을 보면 멀티미디어 검색, 주식 시장 데이터 분석, 센서 기반의 모니터링 등 매우 다양하다[1].

일반적으로 데이터 시퀀스에 대한 분석은 유사 패턴을 찾아내는 유사도 검색을 통해서 이루어진다. DTW (Dynamic Time Warping)은 데이터 시퀀스 간 유사도 검색에 사용되는 대표적인 방법이다. 유클리디언 거리는 데이터 시퀀스를 구성하는 각 요소들을 독립적으로 다뤘다. 이로 인해 길이가 다르거나 샘플링 비율이 다른 데이터 시퀀스간의 거리측정에 적합하지 않았다.

DTW는 유클리디언 거리로 처리할 수 없는 문제를 해결하는 거리 측정 방식이다. DTW는 시간 축에 대해 데이터 시퀀스를 확장하거나 축소 할 수 있도록 하여 데이터간의 최적 거리를 찾아낼 수 있다.

* 본 연구는 지식경제부 산업원천기술개발사업의 일환인 기업 맞춤형 정보전자 패키지 핵심 기술 개발사업 (2009-S-001-010)의 지원을 받아 수행하였음.
접수번호 : #111128-001
접수일자 : 2011년 11월 28일
심사완료일 : 2011년 12월 14일
교신저자 : 송석일, e-mail : sisong@cjnu.ac.kr

하지만, DTW는 그 복잡도로 인해서 대용량 시계열 데이터베이스에서 사용하기에는 한계가 있다. 이러한 문제를 해결하기 위해서 다양한 색인 기법들에 대한 연구가 진행되어왔다[1-7]. 이 방법들을 요약하면, 착오기각 (flase negative)을 발생 시키지 않으면서 시퀀스 데이터베이스를 필터링 할 수 있는 방법을 이용하여 후보 집합을 찾아낸 후에 DTW를 이용해 최종 결과를 찾아낸다.

기존에 제안된 필터링 방법들 중 가장 대표적인 것은 하한 거리 (Lower Bound)와 PAA (Piece-wise Aggregate Approximation) 이다. PAA는 시계열 데이터의 차원을 축소하는 방법이고 하한 거리는 질의 시계열 데이터에 상한 및 하한 범위를 부여하여 거리를 계산하는 방법이다. 하한 거리와 PAA는 같이 사용하여 필터링 효과를 높일 수 있다.

[1] 에서는 기존 방법과는 다르게 DTW 거리를 계산하는 도중에 주어진 값보다 큰 값을 구할 것이 예측될 때 계산을 멈추는 방법을 이용하여 속도를 높인다. [8] 에서는 참조 시퀀스들을 선정하고 참조 시퀀스와 데이터 시퀀스간의 DTW 거리를 구해서, 유클리디언 공간상으로 데이터 시퀀스를 변환한다. 이를 통해 복잡도가 높은 DTW 대신 유클리디언 거리를 측정해서 검색을 수행하는 방법이다.

이 논문에서는 데이터 시퀀스 필터링에 적용할 수 있는 새로운 거리 측정 방식을 제안한다. 제안하는 색인 기법은 기존의 시퀀스 클러스터링 기법 등을 통해 유사한 시퀀스들을 하나의 그룹으로 묶고, 그룹을 대표하는 최소 하한 시퀀스 (MBS, Minimum Bounding Sequence)를 생성한다. 그리고, 질의 시퀀스와 각 그룹을 대표하는 MBS들과의 최소 DTW 거리를 측정하여 질의와 유사한 시퀀스를 포함할 가능성이 있는 그룹을 찾아내고 그룹 내의 시퀀스와의 DTW 거리 측정을 통해서 가장 유사한 시퀀스를 찾아낸다.

이 논문에서 제안하는 것은 유사한 그룹을 대표하는 MBS (Minimum Bounding Sequence) 와 MBS 와 시퀀스 간의 최소 DTW 거리 측정 방식이다. 또한, 이를 이용한 색인 구축 방법 및 질의 처리 방식을 제안한다.

이 논문의 구성은 다음과 같다. 2장에서는 DTW, 하한 거리, PAA 와 같은 관련연구들에 대해서 기술한다. 3장에서는 제안하는 최소 DTW 거리와 이를 이용한 색인 기법에 대해서 기술한다. 4장에서는 제안하는 방법과 하한 거리 기법의 필터링 정도를 실험을 통해 평가하여 제안하는 방법의 우수성을 보인다. 마지막으로 5장에서 결론을 맺는다.

II. 관련 연구

이 장에서는 제안하는 방법의 기반이 되는 DTW와 기존의 색인기법에 사용된 하한거리와 PAA를 기술한다.

1. DTW (Dynamic Time Warping)

두 데이터 시퀀스 Q 와 S의 DTW 거리는 수식 (1)과 같이 계산된다.

$$
\begin{aligned}
&Q = q_1, q_2, \cdots, q_m \\
&S = s_1, s_2, \cdots, s_l \\
&D^p_{DTW}(Q,S) = D^p(q_1, s_1) \\
&\quad + \min \begin{cases} D^p_{DTW}(Q, Rest(S)) \\ D^p_{DTW}(Rest(Q), S) \\ D^p_{DTW}(Rest(Q), Rest(S)) \end{cases}
\end{aligned} \tag{1}
$$

위의 두 시퀀스 Q 와 S 의 길이가 각각 m 과 l 이면 DTW는 O(lm) 의 복잡도를 가진다. DTW의 계산시간을 줄이기 위해 워핑 넓이에 대한 제약을 가하는 방법이 제안된 바도 있다. [그림 1]은 DTW 계산을 위한 행렬이다. 이 그림에서는 시퀀스간에 정렬이 일어날 수 있는 워핑 넓이를 r 로 제한하고 있다.

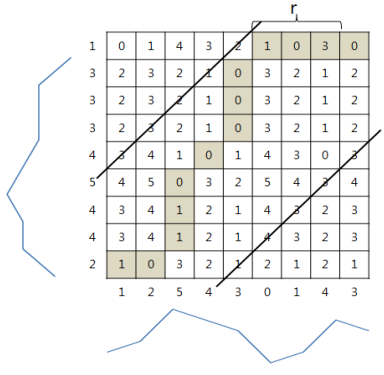

그림 1. DTW Matrix

2. 하한 거리 기법 및 PAA

기존에 제안된 데이터 시퀀스 검색 방법들 중 가장 대표적인 것은 [2]에서 제안하고 있는 하한 거리와 PAA 방법이다. 하한 거리는 질의 시퀀스에 질의 봉투(상한 및 하한 범위)를 설정 하고 질의 봉투와 시퀀스 간의 유클리디언 거리 측정을 통해 후보 시퀀스를 선정하는 것이다. PAA (Piece-wise Aggregate Approximation)는 시퀀스의 차원을 축소하는 기법이다. 하한 거리 측정 방법과 PAA 가 결합이 되어 검색 속도를 향상 시키는 색인 기법이 완성된다.

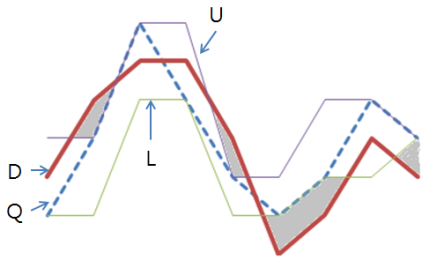

그림 2. 질의 봉투 및 하한 거리

[그림 2]는 [2]에서 제안하는 질의 봉투 및 하한 거리에 대한 예이다. D 는 데이터 시퀀스를 의미하고, Q는 질의 시퀀스이다. U와 L은 각각 Q에 대한 상한 및 하한 범위이다. 그리고, D 와 Q 사이의 거리는 그림 2와 같은 수식에 의해서 계산된다. 즉, 그림 2에서 색이 칠해져 있는 영역의 면적이 바로 두 시퀀스간의 하한 거리가 된다.

언급한 것처럼, PAA 는 시퀀스의 차원을 축소하여 계산 시간을 더욱더 줄이기 위한 차원 축소 방안이다. PAA 로 차원을 축소한 후 하한 거리 기법을 결합하여 필터링 속도를 높인다. 하한 거리에 의해서 선택된 데이터 시퀀스들은 최종 결과가 아니며 DTW를 이용해서 최종 결과를 만들어 내는 정제 과정이 필요하다.

3. 기타 시퀀스 색인 기법

[7] 에서는 새로운 하한 거리 기법을 제안해서 검색 속도를 높이는 방안을 제안하고 있다. 여기에서는 질의 시퀀스와 데이터 시퀀스에 대해 일정 간격으로 구간을 나누고 구간별로 하한, 상한 범위를 부여한다. 그리고, 상한 및 하한 범위를 이용해서 두 시퀀스간의 DTW 거리를 계산한다. 이 거리는 구간별 상한, 하한 범위에 의한 것이므로 근사 값을 갖는다. 이를 통해서 선택된 데이터 시퀀스에 대해서 정제과정을 거쳐 최종 결과를 얻어낸다.

[8] 에서는 기존 방법과는 다른 접근방법을 사용하고 있다. 즉, 근사 거리 측정 방법을 통해 후보 집합을 선정하고 최종 결과를 정제하는 방법을 사용하지 않는다. 이 문헌에서는 대신, 다수의 참조 시퀀스를 선정하고 모든 데이터 시퀀스와 참조 시퀀스들 간의 DTW 거리를 계산하여 각 시퀀스와 참조 시퀀스들 간의 DTW 거리를 그 시퀀스의 특징으로 부여한다. 이와 같은 특징 변환 방식으로 DTW 거리 대신 유클리디언 거리를 이용해서 거리를 측정할 수 있도록 해서 처리 속도를 높인다.

III. 제안하는 색인 기법

이 장에서는 제안하는 최소 DTW 거리 기반의 색인 기법에 대해서 설명한다. 먼저, 최소 DTW 거리에 대한 정의를 하고, 이를 이용한 색인구조 구축 방안, 최근접 질의, 범위질의 처리 방안에 대해서 설명한다.

1. 최소 DTW 거리

최소 DTW 거리는 데이터 시퀀스 그룹에 대한 거리 측정을 위한 것 이다. 본 논문에서는 대용량의 데이터 시퀀스로부터 적절한 클러스터링 방법을 통해 유사한 데이터 시퀀스 그룹을 생성하는 것을 전제로 한다. 시퀀스 클러스터링 방법은 기존에 개발된 어떠한 것도 사용이 가능하다. 생성한 시퀀스 그룹과 질의 시퀀스 사이의 유사도를 측정할 수 있다면 가장 가까운 시퀀스 그룹내의 데이터 시퀀스와 질의 시퀀스를 우선적으로 비교하여 유사한 시퀀스를 빠르게 검색 할 수 있다.

제안하는 최소 DTW 거리는 유사한 시퀀스들의 그룹과 질의 시퀀스 사이의 최소 거리를 구기 위한 거리 측정 방법이다. 시퀀스 그룹과의 거리를 측정하기 위해

서 시퀀스 그룹에 대한 MBS (Minimum Bounding Sequnce)를 정의한다. MBS는 그룹내 데이터 시퀀스들에 대해서 각 차원의 최소 값과 최대 값을 구한 것이다. [그림 3]에서 위쪽은 유사한 시퀀스들을 모아 놓은 시퀀스 그룹이고, 아래쪽은 유사한 시퀀스들에 대한 MBS이다.

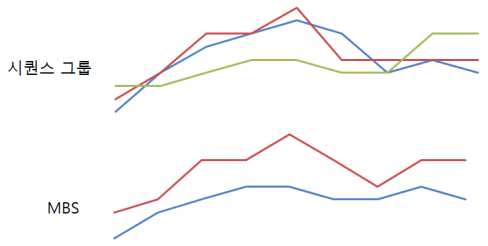

그림 3. MBS 예

MBS 는 식 (3) 과 같이 정의 된다.

$$SG=\{S_0,S_1,\cdots,S_{n-2},S_{n-1}\}$$
$$S_i=\{s_{i,0},s_{i,1},\cdots,s_{i,l-2},s_{i,l-1}\}$$
$$MBS(SG)=\{mbs_0,mbs_1,\cdots,mbs_{l-1}\} \quad (2)$$
$$mbs_i=(l_i,u_i),$$
$$where\, l_i=\min(s_{0,i},s_{1,i},\cdots,s_{n-1,i}) \text{ and}$$
$$u_i=\max(s_{0,i},s_{1,i},\cdots,s_{n-1,i})$$

MBS(SG) 와 질의 시퀀스 Q 와의 최소 DTW 거리는 식 (4) 에 의해서 계산된다. Q 와 MBS 의 I 번째 요소인 q_i와 mbs_i 간의 최소 거리인 $MinDist_i$를 먼저 정의한다. 최소 DTW 거리인 MinDTWDist 는 기본적으로 DTW와 동일하게 계산하며 단지, 질의 시퀀스 Q와 MBS 의 각 요소간 거리를 MinDist 형태로 계산한다는 것이 다르다.

$$Q=\{q_0,q_1,\cdots,q_{l-1}\}$$
$$MinDist_i(q_i,mbs_i)=\begin{cases}|q_i-u_i|, \text{if } q_i>u_i\\ |q_i-l_i|, \text{if } q_i<l_i\\ 0, otherwise\end{cases} \quad (4)$$
$$MinDTWDist(MBS(SG),Q)=$$
$$MinDist_i+$$
$$\min\begin{cases}MinDTWDist(Rest(MBS(SG)),Q)\\ MinDTWDist(MBS(SG),Rest(Q))\\ MinDTWDist(Rest(MBS(SG),Rest(Q))\end{cases}$$

최소 DTW 거리를 시퀀스 그룹들을 필터링 하는데 사용하기 위해서는 어떤 시퀀스 그룹과의 거리는 그룹내의 모든 데이터 시퀀스들과의 거리보다 항상 작거나 같아야 한다.

보조 정리 1. 질의 시퀀스 Q와 시퀀스 그룹 SG={S_0, S_1, ..., S_{l-1}} 의 MBS(SG)와의 거리 MinDTWDist(Q, MBS(SG) 는, SG에 포함되는 각 시퀀스 S_i 와 Q 의 DTW(S_i, Q) 보다 항상 작거나 같다.

증명: DTW를 계산하기 위해서는 두 시퀀스를 구성하는 각 요소들간의 거리를 계산하여 매트릭스를 구성한 후 최소값을 갖는 경로를 찾게 된다. Q 와 S_i (i = 0, 1, ... l-1) 간의 매트릭스를 M(Q, S) 이라 하고, 이 매트릭스의 각 요소를 (m, n) 이라 하자. 또한, Q 와 MBS(SG) 간의 매트릭스를 M(Q, MBS) 라 하고, 각 요소를 (m, n)이라 하자.

이때, M(Q, MBS)의 (m, n)이 M(Q, S) 의 (m, n) 보다 항상 작거나 같다면 MinDTWDist(Q, MBS(SG)) 는 DTW(S_i, Q) 보다 항상 작거나 같다. 제안하는 MinDTWDist 는 MBS의 각 요소와 Q의 각 요소간에 구할 수 있는 거리 중 가장 가까운 거리를 취하므로, M(Q, MBS)의 (m, n)이 M(Q, S) 의 (m, n) 보다 항상 작거나 같다. 따라서, MinDTWDist(Q, MBS(SG)) 는 DTW(S_i, Q) 보다 항상 작거나 같다.

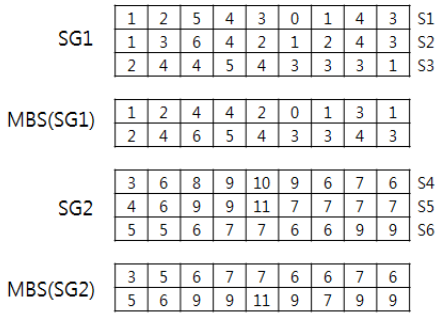

그림 4. MBS 예

[그림 4]와 [그림 5]에서 MinDTWDist를 계산하는 예를 보여주고 있다. [그림 4]에서는 총 6개의 시퀀스 (S_1, S_2, S_3, S_4, S_5, S_6) 가 있으며 이들은 두 개의 시퀀스 그룹 SG1 와 SG2 으로 나뉘어 져 있다. SG1에는 (S_1,

S_2, S_3), SG2 에는 (S_4, S_5, S_6) 이 포함되어 있다. SG1과 SG2에 대한 MBS(SG1) 과 MBS(SG2)가 각각 구해져 있다.

[그림 5]의 위쪽의 매트릭스는 MBS(SG1) 과 질의 시퀀스 Q 와의 MinDTWDist를 구하는 것을 보여주고 있고, 아래쪽 매트릭스는 MBS(SG2) 와 질의 시퀀스 사이의 거리를 구하고 있다. 그림에서처럼 MinDTWDist (Q, MBS(SG1)은 0이고, MinDTWDist (Q, MBS(SG2) 은 31이다. 이로 보아 Q와 보다 유사한 시퀀스들은 SG1에 있으며 Q 와 S_1, S_2, S_3 사이의 DTW 거리가 모두 31 보다 작다면 Q 와 가장 가까운 시퀀스는 SG1에 있는 것을 확정할 수 있다.

Q

2	0	0	2	2	0	0	0	1	0
3	1	0	1	1	0	0	0	0	0
1	0	1	3	3	1	0	0	2	0
0	1	2	4	4	2	0	1	3	1
3	1	0	1	1	0	0	0	0	0
5	3	1	0	0	1	2	2	1	2
5	3	1	0	0	1	2	2	1	2
4	2	0	0	0	0	1	1	0	1
2	0	0	2	2	0	0	0	1	0
	1	2	4	4	2	0	1	3	1
	2	4	6	5	4	3	3	4	3

MBS(SG1)

Q

2	1	3	4	5	5	4	4	5	4
3	0	2	3	4	4	3	3	4	3
1	2	4	5	6	6	5	5	62	5
0	3	5	6	7	7	6	6	7	6
3	0	2	3	4	4	3	3	4	3
5	0	0	1	2	2	1	1	2	1
5	0	0	1	2	2	1	1	2	1
4	0	1	2	3	3	2	2	3	2
2	1	3	4	5	5	4	4	5	4
	3	5	6	7	7	6	6	7	6
	5	6	9	9	11	9	7	9	9

MBS(SG2)

그림 5. MinDTWDist 예

2. 색인구조

MinDTWDist를 이용해서 불필요한 DTW 계산을 줄이기 위해서는 데이터 시퀀스들을 유사한 것들끼리 클러스터링해야 한다. 클러스터링이 잘 될수록 필터링 효과가 높아진다. 유사한 시퀀스들을 하나의 그룹으로 클러스터링 할 수 있는 방법은 K-means 계열의 방법들을 포함해서 매우 다양하다[10]. 하지만, 시퀀스 클러스터링 방법들은 복잡도가 높아서, 데이터 시퀀스가 빈번하게 삽입되거나 기존 삽입된 데이터 시퀀스가 자주 변경되는 동적인 상황에서는 사용하기 어렵다.

하지만, 음악 데이터베이스와 같은 멀티미디어 데이터베이스 등에서는 이미 생성된 데이터 시퀀스에 대한 변경이 거의 없고, 새로운 시퀀스의 발생이 실시간으로 반영되어야 하는 동적인 특성을 가지고 있지 않다. 제안하는 색인 방법은 이와 같은 정적인 환경을 대상으로 하고 있다.

기존의 시퀀스 클러스터링 방법을 이용하여 시퀀스들을 클러스터링하면 다수의 그룹을 얻어 낼 수 있다. 그룹이 생성되면 그룹별로 MBS를 구하여 부가정보와 함께 적절한 자료구조에 저장한다. 기존에 제안된 다양한 다차원 색인 구조를 사용할 수 있지만, 고차원의 데이터 시퀀스에 대한 MBS를 저장하기 위해서는 VA-파일[11]이나 [12] 와 같은 색인 방법이 적절하다고 판단된다.

그룹의 개수가 너무 많을 때는 MBS를 다시 클러스터링해서 그룹들의 그룹을 계층적으로 생성할 수 있다. 그룹들을 다시 클러스터링 해서 생성된 그룹에 대해서도 MBS를 구하고 이를 별도의 VA-파일에 저장한다.

하나의 그룹에 속해 있는 데이터 시퀀스나 MBS는 연속된 디스크 페이지에 저장된다. 되도록 하나의 그룹은 하나의 디스크 페이지에 저장 할수 있도록 한다. 각 그룹은 그룹에 포함된 데이터를 저장하는 페이지 ID 와 그룹의 MBS를 쌍으로 하는 엔트리 (pid, mbs)를 가지고 있다. 이 엔트리들을 VA-파일에 저장하여 검색에 활용한다.

3. 질의 처리 과정

K-NN 질의는 질의 시퀀스 Q와 가장 유사한 시퀀스 K개를 찾아내는 질의이다. K-NN 질의는 다음과 같은 과정을 거쳐서 처리된다. 먼저 VA-파일을 순차 검색하여 질의 시퀀스 Q와 각 그룹의 mbs 와의 MinDTWDist를 구하여 그룹에 대한 엔트리와 함께 우선순위 큐에 넣는다.

그리고, 우선순위 큐에서 MinDTWDist가 가장 작은 그룹의 엔트리를 꺼내서 디스크 페이지로부터 그룹에 속한 데이터 시퀀스를 읽어온다. 읽어온 데이터 시퀀스들과 Q와의 DTW를 거리를 계산하여 가장 가까운 K개의 시퀀스를 결과 집합에 저장한다. 우선순위 큐에서 그 다음으로 가까운 그룹의 엔트리와 거리를 꺼내서 결과 집합의 K 번째 데이터와의 거리와 비교한다. 만일 K번째의 거리가 그룹과의 거리보다 작다면 더 이상 검색을 진행할 필요가 없다.

하지만, 그렇지 않다면, 두 번째 그룹에 속한 시퀀스를 디스크에서 읽어 와서 각 시퀀스와 Q 와의 거리를 계산하고 기존에 결과 집합에 있었던 K 개와 함께 정렬을 해서 다시 K 개의 결과 집합을 찾아낸다. 다음에는, 세 번째로 가까운 그룹 엔트리와 거리를 우선순위 큐에서 읽어온 후, K 번째 시퀀스 거리와 그룹과의 거리를 비교하여 위와 같은 처리를 반복한다.

특정 거리(범위)가 주어지면 질의 시퀀스 Q와 데이터 시퀀스 간의 거리가 범위보다 작은 데이터 시퀀스를 검색하는 질의를 범위질의라 한다. 범위 질의는 상대적으로 처리가 간단하다. VA-파일에서 순차검색을 통해 모든 그룹과 Q 사이의 MinDTWDist를 계산한다. 이들 중, 질의로 주어진 거리보다 작은 그룹들을 걸러내고, 각 그룹에 포함되어 있는 모든 데이터 시퀀스들을 읽어온 후 실제 DTW 거리를 계산하여 최종적으로 주어진 거리보다 거리가 작은 데이터 시퀀스들을 찾아낸다.

IV. 성능 평가

이 논문에서는 제안하는 색인 기법을 검증하기 위해서 기존의 PAA 및 하한 거리를 이용한 색인 기법과 실험을 통해 비교하였다. 성능평가 척도로는 불필요한 DTW 연산의 수행횟수를 선정하였으며, 검색과정에서 수행되는 DTW 연산 횟수를 측정하였다. 실험에 사용된 데이터는 사람이 특정 노래의 일부분을 노래한 음성 파일로부터 특징을 추출한 데이터 시퀀스 이다. 총 200개의 데이터 시퀀스를 실험에 사용하였다. 각 데이터 시퀀스의 차원은 750 이었다.

성능 비교를 위하여 제안하는 방법과 PAA 및 하한 거리를 이용한 색인 방법을 Mac OS 10.7.2, gcc 4.1.1을 기반으로 구현하였다. 이 실험에서는 별도의 클러스터링 방법을 사용하지 않았고 단순히 같은 곡에 대해서 허밍한 노래는 같은 그룹으로 하였다. 또한, 이 실험에서는 VA-파일과 같은 별도의 색인 구조를 이용하지 않았다. 즉, 모든 데이터 시퀀스 및 MBS를 순차검색을 통해서 질의를 처리하도록 하였다. 즉, 모든 MBS를 순차적으로 질의 시퀀스와 비교한 후 가장 작은 거리의 MBS 가 대표하는 시퀀스 그룹내의 데이터 시퀀스들과 순차 비교를 하도록 구현하였다.

질의는 총 1024개의 서로 다른 질의를 수행하였으며, 1024개의 질의중 198개는 데이터 시퀀스베이스에 저장되어 있는 시퀀스를 이용하였다. 질의의 형태는 k-nn 질의를 사용했으며 k의 값은 2, 5, 8, 11, 20 으로 바꿔가며 실험하였다. k 값을 변경할 때 마다 1024개의 질의를 수행하였으며 이때 평균 DTW 연산 횟수를 측정하였다.

[그림 6]은 실험 결과를 보여주고 있다. 그림에서 보는 것처럼 K가 커질수록 제안하는 방법과 기존 방법의 DTW 계산 횟수 차이가 커짐을 알 수 있다. 제안하는 방법의 경우 그룹이 20개 일 때는 항상 20회의 최소 DTW 거리를 계산해야 하며 이는 DTW 계산 횟수에 포함되었다. 만일 그룹을 대표하는 MBS 들을 다시 클러스터링 하여 계층화 한다면 DTW 연산 횟수를 더 줄일 수 있을 것으로 기대한다.

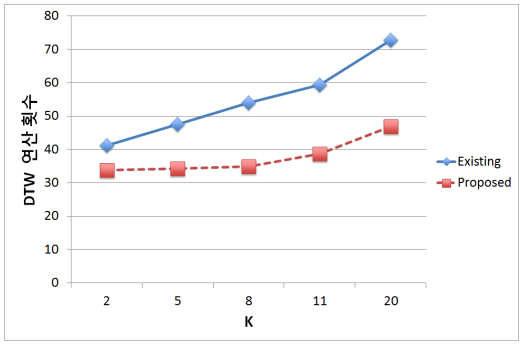

그림 6. K 증가에 따른 DTW 연산 횟수

V. 결론

이 논문에서는 필터링 효과를 높여 불필요한 DTW 연산을 줄이기 위한 시퀀스 색인 기법을 제안하였다. 유사한 데이터 시퀀스가 클러스터링이 되었을 때 생성되는 시퀀스 그룹을 대표하는 MBS를 정의하였다. 그리고, MBS 와 질의 시퀀스와의 거리를 계산하기 위한 최소 DTW 거리를 정의하였다. 또한, MBS와 최소 DTW 거리를 이용하여 K-NN 질의 및 범위 질의를 처리하는 방법을 제안하였다.

제안하는 방법을 검증하기 위해서 기존에 제안된 하한 거리 및 PAA를 이용한 색인 기법과 제안하는 기법을 구현하고 실험을 통해 K-NN 질의를 처리하기 위해 수행하는 DTW 연산의 횟수를 측정하였다. 비록 소규모의 데이터 시퀀스를 기반으로 하는 실험이었지만, 제안하는 방법이 기존방법에 비해 DTW 연산을 적게 함을 볼 수 있었다.

향후에는 제안하는 방법을 바탕으로 계층적으로 색인 구조를 구축하는 방안을 고려할 것이다. 계층적으로 구성하게 되면 보다 높은 필터링 효과를 기대할 수 있다. 또한, 제안하는 방법은 기존에 개발된 다른 방법들[7, 2]과 같이 사용할 수 있는 측면이 있다. 향후 연구에서는 이런 기존방법들과의 결합을 통해 얼마나 성능을 향상 할 수 있는지에 대해서 연구한다. 또한, 실제 응용에서 발생하는 데이터를 이용한 실험을 통해 제안하는 방법의 확장성 및 실세계의 응용에 적합한지를 분석한다.

참 고 문 헌

[1] I. Assent, M. Wichterich, R. Krieger, H. Kremer, and T. Seidl, “Anticipatory DTW for Efficieint Similarity Search in Time Series Databases,” Proceedings of the VLDB Endowment, pp.826-837, 2009.

[2] E. Keogh and C.A. Ratanamahatana, “Exact Indexing of Dynamic Time Warping,” Knowledge and Information Systems, Vol.7, No.3, pp.358-386, 2005.

[3] 김상욱, 박상현, “시퀀스 데이터베이스에서 타임 워핑을 지원하는 효과적인 유사 검색 기법”, 정보과학회논문지:데이터베이스, 제28권, 제4호, pp.643-654, 2001.

[4] 한욱신, 이진수, 문양세, “DTW 거리를 지원하는 범위 서브시퀀스 매칭”, 정보과학회논문지:컴퓨팅의 실제 및 레터, 제14권, 제6호, pp.559-566, 2008.

[5] Y. Zhu and D. Shasha, “Warping Indexes with Envelope Transforms for Query by Humming,” Proceedings of the ACM SIGMOD, pp.181-192, 2003.

[6] S. Salvador and P. Chan, “FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Spaces,” Proceedings of KDD Workshop on Mining Temporal and Sequential Data, pp.70-80, 2004.

[7] Y. Sakurai, M. Yoshikawa, and C. Faloutsos, “FTW : Fast Similiarity Search under the Time Warping Distance,” Proceedings of ACM PODS, pp.326-337, 2005.

[8] V. Athitsos, P. Papapetrou, M. Potamias, G. Kollios, and D. Gunopulos, “Approximate Embedding-based Subsequence Matching of Time Series,” Proceedings of ACM SIGMOD, pp.365-378, 2008.

[9] A. Guttman, “R-trees: A Dynamic Index Structure for Spatial Searching,” Proceedings of ACM SIGMOD, pp.47-57, 1984.

[10] T. Warrenliao, “Clustering of Time Series Data -a Survey,” Pattern Recognition, Vol.38, No.11, pp.1857-1874, 2005.

[11] R. Weber, H. J. Schek, and S. Blott, “A Quantitative Analysis and Performance Study for Similarity Search Methods in High-dimensional Spaces,” Proceedings of VLDB,

pp.194-205, 1998.

[12] 복경수, 허정필, 유재수, “동적 비트 할당을 통한 다차원 벡터 근사 트리”, 한국콘텐츠학회논문지, 제4권, 제3호, pp.81-90, 2004.

저 자 소 개

길 기 정(Ki-jeong Khil) 준회원

- 2011년 2월 : 충주대학교 컴퓨터공학과(공학사)
- 2011년 3월 ~ 현재 : 충주대학교 컴퓨터공학과 석사과정

<관심분야> : 데이터베이스 시스템, 스토리지 시스템

송 석 일(Seokil Song) 정회원

- 2000년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2003년 2월 : 충북대학교 정보통신공학과(공학박사)
- 2003년 8월 ~ 현재 : 충주대학교 컴퓨터공학과 부교수

<관심분야> : 데이터베이스 시스템, 색인구조, 센서 데이터베이스, 클라우드 컴퓨팅

송 재 종(Chai-Jong Song) 정회원

- 1999년 : 원광대학교 전자공학과 학사
- 2001년 : 광운대학교 대학원 전자공학과 석사
- 2001년 ~ 현재 : 전자부품연구원 디지털미디어연구센터 선임연구원

<관심분야> : 멀티미디어 통신, 오디오 신호처리, 디지털 방송 시스템

이 석 필(Seok-Pil Lee) 정회원

- 1990년 : 연세대학교 전기공학과 학사
- 1992년 : 연세대학교 대학원 전기공학과 석사
- 1997년 : 연세대학교 대학원 전기공학과 박사
- 1997년 ~ 2002년 : 대우전자 영상 연구소 선임연구원
- 2002년 ~ 현재 : 전자부품연구원 디지털미디어연구센터 센터장

<관심분야> : 디지털방송통신융합시스템, 멀티미디어 신호처리

장 세 진(Sei-Jin Jang) 정회원

- 1995년 : 경북대학교 전자공학과 졸업(학사)
- 1997년 : 경북대학교 대학원 전자공학과 졸업(석사)
- 1997년 ~ 2002 : 대우전자 영상 연구소 연구원
- 2002년 ~ 현재 : 전자부품연구원 디지털미디어연구센터 책임연구원
- 2010년 ~ 현재 : 전자부품연구원 차세대음향산업지원센터 센터장

<관심분야> : 멀티미디어 통신, 차세대 실감 다채널 오디오, 디지털 방송 시스템

이 종 설(Jong-Seol Lee) 정회원

- 1996년 2월 : 충북대학교 정보통신공학과 학사
- 2001년 2월 : 충북대학교 정보통신공학과 석사
- 2001년 10월 ~ 현재 : 전자부품연구원

<관심분야> : TV-Anytime/Mpeg 시스템, 맞춤형 방송