타임 워핑 하의 효율적인 시계열 서브시퀀스 매칭을 위한 접두어 질의 기법의 확장[1)]

장병철*, 김상욱*, 차재혁*
*한양대학교 정보통신대학원
e-mail:bcchang@ihanyang.ac.kr, {wook, chajh}@hanyang.ac.kr

Extension of the Prefix-Querying Method for Efficient Time-Series Subsequence Matching Under Time Warping

Byoungchol Chang*, Sang-Wook Kim*, Jaehyuk Cha*
*Graduate School of Information and Communications
Hanyang University

요 약

본 논문에서는 타임 워핑 하의 시계열 서브시퀀스 매칭을 처리하는 방법에 대하여 논의한다. 타임 워핑은 시퀀스의 길이가 서로 다른 경우에도 유사한 패턴을 갖는 시퀀스들을 찾을 수 있도록 해 주는 변환이다. 접두어 질의 기법(prefix-querying method)는 착오 기각(false dismissal) 없이 타임 워핑 하의 시계열 서브시퀀스 매칭을 처리하는 인덱스를 이용한 최초의 방식이다. 이 방법은 사용자가 질의를 편리하게 작성하도록 하기 위하여 기본 거리 함수로서 L_∞를 사용한다. 본 논문에서는 L_∞ 대신 타임 워핑 하의 시계열 서브시퀀스 매칭에서 기본 거리 함수로서 가장 널리 사용되는 L_1을 적용할 수 있도록 접두어 질의를 확장한다. 또한, 제안된 기법으로 타임 워핑 하의 시계열 서브시퀀스 매칭을 수행하는 경우 착오 기각이 발생하지 않음을 이론적으로 증명한다. 다양한 실험을 통한 성능 평가를 통하여 본 연구에서 제시하는 기법의 우수성을 검증한다. 실험 결과에 의하면, 제안된 기법은 가장 좋은 성능을 보이는 기존의 기법과 비교하여 매우 뛰어난 성능 개선 효과를 보이는 것으로 나타났다.

1.서론

시계열 데이터베이스(time-series database)란 객체의 변화되는 값들의 연속으로 구성된 데이터 시퀀스(data sequence)들의 집합이다[Agr93]. 시퀀스 매칭(sequence matching)이란 주어진 질의 시퀀스(query sequence)와 변화의 패턴이 유사한 시퀀스들을 시계열 데이터베이스로부터 찾아내는 데이터 마이닝(data mining) 및 데이터 웨어하우징(data warehousing) 분야의 중요한 연산이다[Agr93].

시퀀스 매칭에 관한 대부분의 기존 연구에서는 길이 n의 시퀀스를 n 차원 공간상의 한 점으로 간주한다. 또한, 길이 n인 서로 다른 두 시퀀스 $X(=[x_1, x_2, \ldots, x_n])$와 $Y(=[y_1, y_2, \ldots, y_n])$ 간의 유사한 정도를 측정하는 척도로서 아래의 식과 같이 정의되는 거리 함수 $L_p(X,Y)$를 널리 사용한다. L_1은 맨하탄 거리(Manhattan distance), L_2는 유클리드 거리(Euclidean distance), L_∞은 대응되는 각 요소 값 쌍의 거리 중 최대 거리를 의미한다. 응용에서 주어진 허용치 ε보다 작거나 같은 $L_p(X,Y)$를 갖는 임의의 두 시퀀스 X, Y를 유사하다고 간주한다[Agr93].

$$L_p(X,Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

L_p 거리 함수만을 이용한 시퀀스 매칭을 통해서는 사용자가 원하는 시퀀스들을 검색하지 못하는 경우가 빈번하게 발생한다. 따라서 응용 분야에 적합한 유사 모델(similarity model)을 적절하게 정의할 수 있도록 변환(transform)을 지원하기도 한다.

여러 가지 변환 방법 중 타임 워핑은 시퀀스내의 각 요소 값을 임의의 수만큼 반복시키는 것을 허용하는 변환이다[Yi98]. 이는 비교하는 두 시퀀스의 길이가 다른 경우의 유사도 측정에도 유용하게 사용할 수 있다[Kim01]. 타임 워핑 후의 두 시퀀스들 간의 거리를 타임 워핑 거리(time warping distance)라 한다. 두 시퀀스 S와 Q간의 타임 워핑 거리(time warping distance) D_{tw}는 다음과 같이 재귀적으로 정의된다[Yi98][Par00][Kim01]:

정의 1:

(1) $D_{tw}(<>, <>) = 0$,
(2) $D_{tw}(S, <>) = D_{tw}(<>, Q) = \infty$,
(3) $D_{tw}(S, Q) = (|L_p(First(S), First(Q))|^p + |\min(D_{tw}(S, Rest(Q)), D_{tw}(Rest(S), Q), D_{tw}(Rest(S), Rest(Q)))|^p)^{1/p}$

여기서, First(S)는 S의 첫 번째 요소 s_1을 의미하며, Rest(S)는 s_1을 제외한 S의 나머지 요소들로 구성되는 시퀀스를 의미한다. <>은 요소가 존재하지 않는 널 시퀀스(null sequence)를 의미한다. min은 세 개의 인자들 중 가장 작은 값을 가지는 것을 취하는 함수이다. L_p는 응용에서 적합한 것을 선택하여 사용할 수 있으나, 현재 맨해튼 거리(Manhattan distance) L_1을 기반으로 하는 타임 워핑 거리가 가장 널리 사용되고 있다.

타임 워핑 하의 시퀀스 매칭의 처리를 위해 수행된 최근 연구 기법들은[Ber96][Yi98][Par00][Kim01][Par01] 크게 여과 단계(filtering step)와 후처리 단계(post processing step)로 구성된다. 여과 단계는 주어진 질의 시퀀스와 전혀 유사하지 않는 시퀀스들을 미리 제거함으로써 최종 결과에 포함될 가능성이 매우 높은 시퀀스들로만 구성되는 후보 집합(candidate set)을 구성하는 단계이다. 질의 시퀀스와 실제로 유사한 시퀀스를 필터링 단계에서 후보 집합 내에 포함시키지 못하는 현상을 착오 기각(false dismissal)이라 한다. 반면, 질의 시퀀스와 유사하지 않은 일부의 시퀀스를 필터링 단계에서 후보 집합 내에 포함시키는 현상을 착오 채택(false alarm)이라 한다. 후처리 단계는 후보 집합에 속하는 각 시퀀스를 디스크로부터 액세스하여 이것이 질의 시퀀스와 실제로 유사한가의 여부를 판단함으로써 착오 채택을 제거하는 단계이다. 참고 문헌 [Ber96]에서는 여과 단계 없이 모든 시퀀스들 각각에 대하여 질의 시퀀스와의 타임 워핑 거리를 동적 프로그래밍

1) 본 논문은 정보통신부 IT 연구센터(강원대학교 미디어서비스기술 연구센터), 한양대학교 교내 연구비(HY-2003), 제주대학교 IT연구센터(텔리매틱스 요소 기술 연구)의 부분적인 지원을 받았습니다.

(dynamic programming)을 이용하여 계산함으로써 타임 워핑 하의 시퀀스 매칭을 처리하는 Naïve-Scan 기법[Kim01]을 제안하였다. 여과 단계 없이 모든 시퀀스들을 후보로 고려하는 Naïve-Scan은 그 처리 성능이 떨어지므로, 참고 문헌 [Yi98]과 [Par00]에서는 여과 단계를 채택하는 LB-Scan과 ST-Filter기법들을 제안하였다. LB-Scan[Yi98]은 별도의 자료 구조 없이 모든 시퀀스들을 대상으로 여과 단계를 신속하게 수행하고, 이 결과 반환되는 후보 시퀀스들만을 대상으로 후처리 단계를 수행한다. 반면, ST-Filter[Par00]는 접미어 트리(suffix tree)라는 별도의 자료 구조를 이용하여 여과 단계를 수행하고, 이 결과 반환되는 후보 시퀀스들만을 대상으로 후처리 단계를 수행한다. LB-Scan과 ST-Filter 모두 후처리 단계에서는 Naïve-Scan에서와 같이 동적 프로그래밍을 이용하여 타임 워핑 거리를 계산한다. 시퀀스 매칭은 전체 매칭(whole matching)과 서브시퀀스 매칭(subsequence matching)으로 구분된다[Fal94]. 서브시퀀스 매칭은 전체 매칭을 일반화한 것이므로[Fal94][Yi98] 실제 응용 분야에서 널리 사용된다.

저자가 참여한 참고 문헌 [Kim01]에서는 인덱스 기반의 타임 워핑하의 시퀀스 매칭 기법을 제안한 바 있고, 참고 문헌 [Par01]에서는 이를 확장한 서브시퀀스 매칭 기법인 접두어 질의 기법을 제안한 바 있다. 이 두 기법들은 사용자의 질의 편리성을 위하여 L_∞을 기본 거리 함수로 사용한다. 그러나 국내외 학술회의 또는 학술지를 통한 논의 및 심사과정에서 본 저자들은 L_1을 위한 인덱스 기반 타임 워핑하의 시퀀스 매칭 기법의 확장에 대한 향후 연구에 대하여 요청을 받은 바 있다[Fal01]. 본 논문은 이러한 요청에 대한 답으로서 작성되었다.

본 논문에서는 타임 워핑을 지원하는 인덱스 기반 서브시퀀스 매칭 기법에 관하여 논의하고자 한다. 기존 접두어 질의 기법[Par01]이 거리 함수 L_1을 적용하는 경우에도 올바르게 동작 하는가의 여부는 현재까지 검증된 바 없다. 본 논문에서는 L_1을 적용할 수 있도록 접두어 질의 기법을 확장하는 방안을 제시한다. 또한 확장된 기법이 착오 기각(false dismissal)없이 검색 대상이 되는 모든 서브시퀀스들을 올바르게 검색한다는 것을 이론적으로 증명한다. 제안된 기법의 우수성을 규명하기 위하여 상기 언급한 기존의 기법들과 실험을 통한 성능 평가를 수행한다.

2. 접두어 질의의 확장

참고 문헌 [Kim01]에서 LB-Filter 기법의 D_{tw_lb}가 D_{tw}의 하한 함수인 동시에 유사 검색에서 사용하는 거리 함수 L_∞에 대해 삼각 부등식을 만족함을 보임으로써 착오 기각이 발생하지 않음을 증명하였다. 본 연구에서는 이를 확장하여 거리 함수 L_1에 대해서도 접두어 질의 기법을 적용할 수 있음을 증명하고자 한다.

정의 2:
$D_{tw\_lb}(S, Q) = L_1(\text{Feature}(S), \text{Feature}(Q))$ 여기서
Feature(S)=<First(S), Last(S), Greatest(S), Smallest(S)>,
Feature(Q)=<First(Q), Last(Q), Greatest(Q), Smallest(Q)> 이다.

다음에는 정리 1과 정리 2를 이용하여 함수 D_{tw_lb}가 타임 워핑 거리 D_{tw}의 하한 함수인 동시에 삼각 부등식을 만족함을 보이고자 한다. 정리 1의 증명을 위하여 다음의 보조 정리 1과 보조 정리 2 및 가정 1을 이용한다. 가정 1은 본 논문의 정리 3과 그것의 따름 정리 2까지 적용된다. 정리 5에서는 가정 1이 성립하지 않는 경우에 대한 해결 방안을 제시 한다.

가정 1:
데이터 시퀀스 S와 질의 시퀀스 Q의 워핑된 시퀀스를 S′, Q′라 하면, S′ = <s_1', s_2', ..., s_k'>, Q′ = <q_1', q_2', ..., q_k'>에서 s_1', s_k', q_1', q_k'은 각 시퀀스 내에서 최대값 또는 최소값이 아니다.

보조 정리 1:
임의의 두 시퀀스 S = <s_1, s_2, ..., s_n>, Q = <q_1, q_2, ..., q_m>에 대하여 다음이 항상 성립한다.

$D_{tw}(S, Q) \geq L_1(\langle First(S), Last(S)\rangle, \langle First(Q), Last(Q)\rangle)$

증명: 참고 문헌 [Cha05] 참조.

보조 정리 2:
임의의 두 시퀀스 S = <s_1, s_2, ..., s_n>, Q = <q_1, q_2, ..., q_m>에 대하여 다음이 항상 성립한다.

$D_{tw}(S, Q) \geq$
$L_1(\langle Greatest(S), Smallest(S)\rangle, \langle Greatest(Q), Smallest(Q)\rangle)$

증명: 참고 문헌 [Cha05] 참조.

정리 1:
임의의 두 시퀀스 S = <s_1, s_2, ..., s_n>, Q = <q_1, q_2, ..., q_m>에 대하여 다음이 항상 성립한다.

$$D_{tw}(S, Q) \geq D_{tw\_lb}(S, Q)$$

정리 1을 이용하여 다음의 따름 정리 1을 쉽게 유도해 낼 수 있다.

따름 정리 1:
임의의 두 시퀀스 S = <s_1, s_2, ..., s_n>, Q = <q_1, q_2, ..., q_m>와 임의의 값 ε에 대하여 다음이 항상 성립한다.

$$D_{tw}(S, Q) \leq \varepsilon \Rightarrow D_{tw\_lb}(S, Q) \leq \varepsilon$$

정리 2:
임의의 세 시퀀스 X, Y, Z에 대하여 다음이 항상 성립한다.

$$D_{tw\_lb}(X, Z) \leq D_{tw\_lb}(X, Y) + D_{tw\_lb}(Y, Z)$$

증명: 참고 문헌 [Cha05] 참조.

정리 3:
임의의 두 시퀀스 s, q, 그리고 임의의 양수 w($1 \leq w \leq |s|$)에 대하여, 만일 s와 q의 타임 워핑 거리가 ε이내이면, s의 접두어 s[1:w]와의 타임 워핑 거리가 ε이내인 q의 접두어가 반드시 존재한다. 즉, 아래의 공식이 성립한다.

$$D_{tw}(s, q) \leq \varepsilon \Rightarrow (\exists x)(D_{tw}(s[1:w], q[1:x]) \leq \varepsilon)$$

증명: 참고 문헌 [Cha05] 참조.

LB-Filter에서 사용하였던 거리 함수 D_{tw_lb}는 타임 워핑 거리 D_{tw}의 하한 함수이므로, 정리 3으로부터 다음과 같은 따름 정리 2을 쉽게 유도할 수 있다.

따름 정리 2:
임의의 두 시퀀스 s, q, 그리고 임의의 양수 w($1 \leq w \leq |s|$)에 대하여, 다음의 식이 항상 성립한다.

$$D_{tw}(s, q) \leq \varepsilon \Rightarrow (\exists x)\ (D_{tw\_lb}(s[1:w], q[1:x]) \leq \varepsilon),$$
여기서 $1 \leq x \leq |q|$

특성 벡터 내 다중 역할 요소 문제

가정 1의 증명에서 워핑된 시퀀스 S′ = <s_1', s_2', ..., s_k'>, Q′ = <q_1', q_2', ..., q_k'>에서 s_1', s_k', q_1', q_k'이 최대값 또는 최소값이 아니라고 전제하였다. 그러나 흔하지 않은 경우이지만 실제 응용에서는 가정 1이 만족되지 않는 경우가 발생될 수 있다. 여기서는 이 문제에 대하여 고찰하고, 해결 방안을 제시한다.

워핑된 시퀀스 S′ = <s_1', s_2', ..., s_k'>에서 s_1', s_k'이 최대값 또는 최소값이 된다는 것은 Feature(S) = <First(S), Last(S), Greatest(S), Smallest(S)> 에서 Fisrt(S) = Greatest(S) or Smallest(S) 이거나 Last(S) = Greatest(S) or Smallest(S)가 된다는 것을 의미한다. 이는 s_1', s_k'와 동일한 값이 s_2', s_3', ..., s_{k-1}'에 존재하는 경우와 s_1', s_k'와 동일한 값이 s_2', s_3', ..., s_{k-1}'에 존재하지 않는 경우로 구분하여 생각할 수 있다. 전자는 문제가 발생하지 않느나[Cha05] 후자의 경우는 시퀀스에서 특성 벡터를 구성하면 둘 이상의 특성으로 표현되는 요소가 포함된 특성 벡터가 만들어질 수 있으며, 이 요소를 다중 역할 요소(multi-role element)라 정의 한다. 특성 벡터 내에 다중 역할 요소가 있다는 것은 특성 벡터의 두 특성이 워핑된 시퀀스내의 같은 한 요소로부터 추출 되었다는 것을 의미한다. 따라서 Greatest(S) 혹은 Smallest(S)가 <s_2', s_3', ..., s_{k-1}'>에 존재하지 않으므로 정리 1을 보장할 수 없다.

정리 4:
접두어 질의가 L_1에 대하여 삼각 부등식을 만족하여 착오

기각이 발생하지 않으려면 임의의 시퀀스 S의 특성 벡터 Feature(S) = <First(S), Last(S), Greatest(S), Smallest(S)>를 구성하는 네 요소들은 워핑된 시퀀스 $S' = <s_1', s_2', s_3', ..., s_{k-1}', s_k'>$의 서로 다른 요소이어야 한다.
증명: 참고 문헌 [Cha05] 참조.

정리 5:
데이터 시퀀스 S 또는 질의 시퀀스 Q의 특성 벡터에 다중 역할 요소가 포함되어 있을 때, 2배의 ε값을 기준으로 질의를 수행하면 착오 기각이 발생하지 않는다.
증명: 참고 문헌 [Cha05] 참조.

2배 확장된 ε로 질의하면 특성 벡터 내에 다중 역할 요소가 존재하여도 착오 기각이 발생하지 않아 접두어 질의 기법에 L_1을 적용할 수 있다. 그러나 ε의 값이 2배가 된 것에 비례하여 착오 해답(false alarm)이 증가하게 된다. 본 논문에서는 이배수 ε 질의 시 발생되는 착오 해답을 줄여 성능을 높이기 위한 방안으로 초기 데이터 시퀀스에서 특성 벡터를 추출하여 트리를 구성할 때, 두 종류의 트리를 구성하는 방법을 제안한다. 데이터 시퀀스에서 특성 벡터를 다중 역할 요소 없이 추출할 수 있는 시퀀스로 구성된 트리 T_d와 그렇지 못한 트리 T_i로 두 종류의 트리를 구성한다. 단, 트리를 구성하는 시점에는 워핑된 시퀀스가 아니므로 다중 역할 요소의 유무를 정확히 알 수 없다. 따라서 시퀀스 내에 First값과 Last값이 최대값 혹은 최소값인 시퀀스를 모두 T_i로 구성한다. 이는 질의 시퀀스에 대해서도 동일하게 적용된다. 질의 시퀀스가 들어오면 T_d에는 정상적인 ε 질의를 시행하고 T_i에는 이배수 ε 질의를 시행한다. 만일 질의 시퀀스의 특성 벡터가 다중 역할 요소를 가지는 경우에는 T_d와 T_i 모두에 이배수 ε 질의 기법을 사용하여 질의 한다. 이 경우에는 착오 해답이 증가할 수 있다. 그러나 길이가 길고 그 값이 다양한 실제 시퀀스 데이터에서 특성 벡터를 구성하는 경우 다중 역할 요소가 존재하는 경우는 매우 드물게 나타난다.

3. 성능 평가

본 장에서는 접두어 질의 기법의 성능을 서론에서 제시한 Naïve-Scan, LB-Scan, ST-Filter와 비교 분석하고자 한다.

3.1 실험 환경

본 실험에서는 성능 분석을 위하여 한국의 실제 주식 데이터로서 길이가 300인 620개의 데이터 시퀀스로 구성된 데이터베이스 K_Stock_Data를 사용하였다.

질의 시퀀스 Q는 데이터베이스에서 선택한 시퀀스로부터 길이가 Len(Q)인 임의의 서브시퀀스를 선택하여 그대로 사용하는 방법으로 생성하였다. 질의 구성 시에는 질의 선택률(query selectivity)을 아래의 식과 같이 정의하고, 각 질의에 대하여 원하는 선택률을 만족하도록 허용치 ε을 조정하였다.

$$\text{선택률} = \frac{Q\text{와 } \varepsilon - \text{매치하는 모든 서브시퀀스들의 수}}{Q\text{와 매치 가능한 길이를 가진 모든 데이터 서브시퀀스들의 수}}$$

성능 평가를 위한 하드웨어 플랫폼은 1.7GHz Pentium IV CPU와 1.2GB의 주기억장치가 장착된 PC이며, 소프트웨어 플랫폼은 Linux kernel version 2.4.18 및 Glibc 2.2.4이다. 실험 중 다른 프로세스들과의 상호 간섭을 방지하기 위하여 운영체제를 단일 사용자 모드로 설정해 모든 사용자 프로세스들을 제거한 상황에서 실험하였다. 또한 ST-Filter를 위한 도메인 분류 방법으로서 최대 엔트로피 기법(maximum entropy method)을 이용하여 ST-Filter가 50개의 구간을 갖도록 하였다.

3.2 실험 결과 및 분석

실험 1에서는 선택률을 7.1×10^{-8}, 2.13×10^{-7}, 3.55×10^{-7}, 4.97×10^{-7}로 변화하면서 접두어 질의 기법과 여과 단계를 거치는 기존 기법들의 성능을 비교하였다. 사용된 질의 시퀀스의 길이는 110이며, 각각 2, 6, 10, 14개를 최종 결과로 반환하였다. 그림 3.1은 실험 결과를 나타낸 것이다.

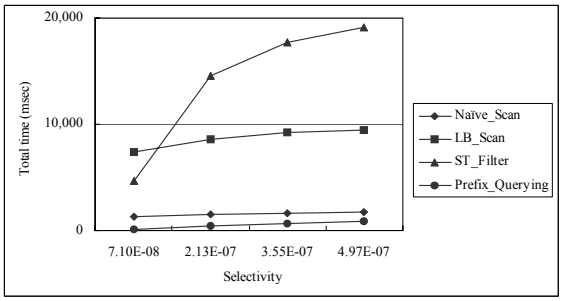

그림 3.1. 선택률 변화에 따른 성능 결과.

선택률이 증가함에 따라 모든 기법에서 검색 시간이 증가하였다. ST-Filter 기법의 경우 그 증가폭이 크게 나타나고 있다. 이는 L_1의 경우 접미어 트리에서 후보를 검색할 때 거리 함수 테이블을 구성하는데 요구되는 CPU 연산 양이 ε 값에 많은 영향을 받기 때문이다. LB-Scan 기법의 경우 그 특성상 후보 시퀀스를 구성 할 때 모든 데이터 시퀀스를 액세스해야 하므로 비교적 좋지 않는 성능을 나타내고 있다.

본 논문의 실험에서 주목할 점은 Naïve-Scan 기법이 모든 경우에서 LB-Scan나 ST-Filter 기법 보다 성능이 우수한 것으로 나타난 것이다. 이것은 본 논문의 실험에서 CPU 처리 과정을 최적화 한 개선된 Naïve-Scan[Kim05]을 사용하였기 때문이다. 본 논문에서 사용된 Naïve-Scan 방식은 질의 시퀀스와 서브시퀀스들 간의 타임 워핑 거리를 계산하는 과정에서 발생되는 많은 중복 작업을 사전에 제거하는 방식으로 CPU 성능을 극대화하는 방식이다. 따라서 이 Naïve-Scan 방식을 이용하면 기존에 CPU 처리에서 발생하는 성능 병목 현상이 줄어들어 본 실험과 같이 우수한 성능을 보이는 것이다. 이러한 성능 역전 현상은 이후의 모든 실험에서 일관되게 나타났으며, 이는 참고 문헌 [Kim05]의 실험 결과와도 부합되는 것이다. 성능 역전 현상에 대해서는 이후의 실험에서 별도로 언급하지 않는다. 접두어 질의 기법은 다른 기법들과 비교하여 월등한 성능을 보였으며, 기존 기법 중 가장 좋은 성능을 보인 Naïve-Scan에 비해 최대 10.7배 더 좋은 성능을 보였다. 이것은 인덱스를 이용함으로써 여과 단계의 성능을 크게 개선할 수 있기 때문이다.

시계열 데이터에 타임 워핑 기법을 적용 할 때, 경우에 따라 하나의 시퀀스 요소가 너무 많이 반복되지 않도록 한 요소가 반복될 수 있는 최대의 수를 정해 놓는다. 이렇게 정해 놓은 수를 maxWarpRatio라고 한다[Ber96]. 실험 2에서는 maxWarpRatio를 3, 6, 9, 12로 변화시키면서 접두어 질의 기법과 기존 기법들의 성능을 비교하였다. 사용된 질의 시퀀스의 길이는 110이고 선택률은 2.13×10^{-7}이다. 그림 3.2는 실험 결과를 나타낸 것이다.

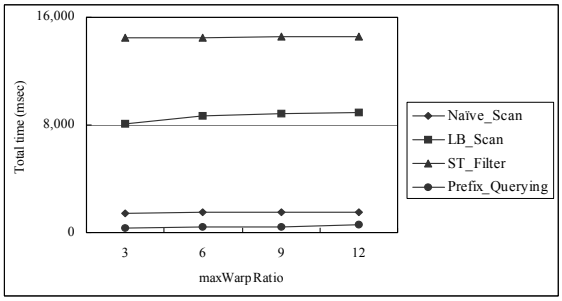

그림 3.2. maxWarpRatio 변화에 따른 성능 결과.

모든 기법에서 maxWarpRatio가 증가함에 따라 여과 단계에서 결과로 반환되는 서브시퀀스들의 수와 후보 서브시퀀스의 수가 많아져서 소요 시간이 조금씩 증가하는 것을 볼 수 있다. 접두어 질의 기법의 경우 최소 질의 시퀀스 길이(minQLen)[Par01]를 maxWarpRatio로 나눈 값을 윈도우 크기

$w = \lceil \frac{minQLen}{maxWarpRatio} \rceil$로 사용하므로 maxWarpRatio가 커지면 접두어의 수가 증가하고 이로 인해 검색된 후보 서브시퀀스의 수가 증가하고 있어 후처리 시간이 점차 증가하는 것으로 나타났다. 이 실험에서 접두어 질의 기법은 나머지 기법 중 가장 좋은 성능을 보인 Naïve-Scan 기법에 비해 최대 4.4배 더 좋은 성능을 보이고 있다.

실험 3에서는 minQLen를 10, 40, 70, 100으로 변화시키면서 접두어 질의 기법과 기존 기법들의 성능을 비교하였다. 사용된 질의 시퀀스의 길이는 110이고, 선택률은 2.13×10^{-7}이다. 그림 3.3은 실험 결과를 나타낸 것이다.

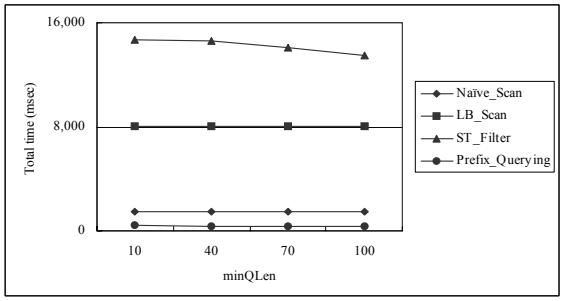

그림 3.3. minQLen 변화에 따른 성능 결과.

ST-Filter 기법은 minQLen가 증가할수록 접미어 트리에 포함되는 접미어의 수가 감소하므로 여과 과정의 계산량이 줄어들어 전체 검색 시간이 줄어드는 것으로 나타났다. 다른 기법들에서는 minQLen의 값에 거의 영향을 받지 않고 대부분 일정한 검색 성능을 보였다. 이 실험에서도 역시 접두어 질의 기법은 나머지 기법 중 가장 좋은 성능을 보인 Naïve-Scan 기법에 비해 최대 3.8배 더 좋은 성능을 보이고 있다.

실험 4에서는 질의 시퀀스의 길이를 80, 140, 200, 260으로 변화시키면서 접두어 질의 기법과 기존 기법들의 성능을 비교하였다. 사용된 질의의 선택률은 2.13×10-7이다. 그림 3.4은 실험 결과를 나타낸 것이다.

질의 시퀀스의 길이가 증가함에 따라 거리 계산 비용이 증가하게 되고, 이로 인해 여과 단계와 후처리 단계에서 소요되는 시간은 모두 그만큼 증가하게 된다. ST-Filter 기법과 LB-Scan 기법은 매우 급격한 검색 시간의 증가를 보여준다. 반면, Naïve-Scan과 접두어 질의 기법은 질의 시퀀스 증가에 따라 매우 완만한 검색 시간의 증가를 보였다. 이 실험에서 역시 접두어 질의 기법은 가장 우수한 성능을 보였으며, Naïve-Scan과 비교하여 최대 3.6배의 성능 개선을 나타냈다.

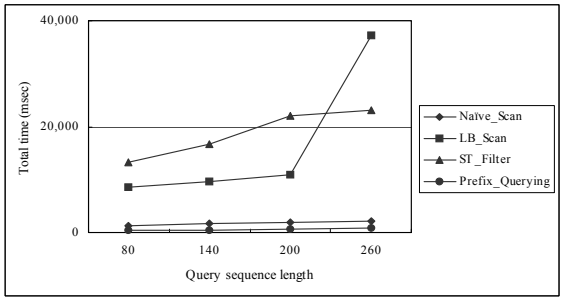

그림 3.4. 질의 시퀀스 길이 변화에 따른 성능 결과.

4. 결론

접두어 질의 기법은 착오 기각 없이 타임 워핑 하의 시계열 서브시퀀스 매칭을 처리하는 인덱스를 이용한 최초의 방식이다. 이 기법은 사용자가 질의를 편리하게 작성하도록 하기 위하여 기본 거리 함수로서 L_∞를 사용한다. 본 논문에서는 L_∞ 대신 타임 워핑 하의 시계열 서브시퀀스 매칭에서 기본 거리 함수로서 가장 널리 사용되는 L_1을 적용할 수 있도록 접두어 질의를 확장하는 방안에 대하여 논의하였다. 본 논문의 주요 공헌은 아래와 같이 요약될 수 있다.

- L_∞ 대신 L_1을 기본 거리 함수로서 적용할 수 있도록 접두어 질의 기법을 확장하였다.
- 확장된 접두어 질의 기법을 이용한 타임 워핑 하의 시계열 서브시퀀스 매칭에서 착오 기각이 발생하지 않음을 이론적으로 규명하였다.
- 다양한 실험에 의한 기존 기법들과의 성능 비교를 통하여 확장된 접두어 질의 기법의 우수성을 규명하였다. 실험 결과에 의하면, 확장된 접두어 질의 기법은 기존의 가장 좋은 성능을 보이는 기법과 비교하여 매우 극적인 성능 개선 효과를 보이는 것으로 나타났다.

대용량의 데이터베이스의 경우 확장된 접두어 질의 기법에서도 여과 단계 후에 발생되는 후보 서브시퀀스들의 개수가 증가되면 성능이 저하된다. 따라서 여과 단계에서 후보 서브시퀀스들의 수를 좀더 줄일 수 있는 방안이 요구된다. 이를 위하여 향후 연구로서 타임 워핑 거리에 좀더 가까운 값을 반환하는 하한 함수를 고안하는 것을 고려하고 있다.

참고 문헌

[Agr93] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient Similarity Search in Sequence Databases," In Proc. Int'l. Conf. on Foundations of Data Organization and Algorithms, FODO, pp. 69-84, Oct. 1993.

[Ber96] D. J. Berndt and J. Clifford, "Finding Patterns in Time Series: A Dynamic Programming Approach," Advances in Knowledge Discovery and Data Mining, pp. 229-248, 1996.

[Cha05] B. C. Chang, S. W. Kim and J. H. Cha, "Extension of the Prefix-Querying Method for Efficient Time-Series Subsequence Matching Under Time Warping", 2005. (Unpublished Manuscript)

[Fal94] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-series Databases," In Proc. Int'l. Conf. on Management of Data, ACM SIGMOD, pp. 419-429, May 1994.

[Fal01] C. Faloutsos, private communication, 2001.

[Kim01] S. W. Kim, S. H. Park, and W. W. Chu, "An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," In Proc. Int'l. Conf. on Data Engineering, IEEE ICDE, pp. 607-614, 2001.

[Kim05] Man-Soon Kim, Sang-Wook Kim, and Mi-Young Shin, "Optimization of Subsequence Matching Under Time Warping in Time-Series Databases," ACM Symp. on Applied Computing, pp. 581-586 Apr. 2005.

[Par00] S. H. Park et al., "Efficient Searches for Similar Subsequences of Difference Lengths in Sequence Databases," In Proc. Int'l. Conf. on Data Engineering, IEEE ICDE, pp. 23-32, 2000.

[Par01] S. H. Park, S. W. Kim, J. S. Cho, and S. Padmanabhan, "Prefix-Querying: An Approach for Effective Subsequence Matching Under Time Warping in Sequence Databases," In Proc. ACM Intl. Conf. on Information and Knowledge Management, ACM CIKM, pp. 255-262, 2001.

[Yi98] B. K. Yi, H. V. Jagadish, and C. Faloutsos, "Efficient Retrieval of Similar Time Sequences Under Time Warping," In Proc. Int'l. Conf. on Data Engineering, IEEE ICDE, pp. 201-208, 1998.