다수의 MBR을 이용한 시계열 서브시퀀스 매칭 연구

임선영*, 박영호*
*숙명여자대학교 멀티미디어과학과
e-mail:{sunnyihm, yhpark}@sookmyung.ac.kr

A Study on Time-Series Subsequence Matching using Multi MBRs

Sun-Young Ihm*, Young-Ho Park*
*Dept of Multimedia Science, Sookmyung Women's University

요 약

시계열 데이타는 일정 시간 간격으로 측정한 값의 시퀀스를 뜻하는데, 사용자에 의해 주어진 질의 시퀀스와 유사한 데이타 시퀀스를 검색하는 방법을 유사 시퀀스 매칭이라고 한다. 본 논문에서는 유사 시퀀스 매칭 시, 질의 시퀀스로 MBR을 구성할 때 한 개의 MBR이 아닌 다수의 MBR로 구성하는 방법을 제안하였다. 다수의 MBR로 구성하여 질의 처리를 하면 질의 시퀀스의 길이가 길 경우 적은 비용으로 질의 처리를 수행할 수 있다.

1. 서론

시계열 데이타(time-series data)는 데이타 마이닝과 데이타 웨어하우스와 같은 데이타베이스의 새로운 응용 분야에서 그 중요성을 더해가고 있다[3]. 시계열은 일정 시간 간격으로 배치된 데이타들의 수열을 말하므로, 즉, 시계열 데이타란 일정 시간 간격으로 측정한 값의 시퀀스를 말한다. 시계열 데이타의 대표적인 예로는 주식 데이타, 환율 변동 데이타, 노래의 음정 변동 데이타 등이 있다. 시계열 데이타베이스에 저장된 시계열 데이타를 데이타 시퀀스(data sequence)라 부르며, 사용자에 의해 주어진 질의 시퀀스(query sequence)와 유사한 데이타 시퀀스를 검색하는 방법을 유사 시퀀스 매칭(similar sequence matching)이라고 한다[1,4]. 컴퓨터의 계산 및 저장 능력이 발전함에 따라 많은 양의 시계열 데이타를 활용하고자 하는 연구가 활발하게 이루어져 왔으며, 특히 시계열 데이타에 대한 유사 시퀀스 매칭은 데이타 마이닝의 중요한 분야로 자리 잡고 있다[2,4].

본 논문에서는 두 시퀀스 사이의 거리를 유클리디안 거리 D(X,Y)로 구한 후, 시퀀스 매칭에 관한 정의는 Dual Match[4]에서 정의한 방법을 사용한다. 두 시퀀스 사이의 거리인 D(X,Y)가 사용자가 제시한 허용치(tolerance)인 ε 이하이면 시퀀스 X와 시퀀스 Y는 유사(similar)하다고 한다[4]. 시퀀스 X와 Y 사이의 거리가 ε 이하이면 X와 Y는 ε-매치(ε-match)한다고 정의하고, 길이 n인 시퀀스들간의 유클리디안 거리를 계산하는 연산을 n 차원 거리 계산(n-dimensional distance computation)이라 정의한다[4].

유사 시퀀스 매칭에서 질의 시퀀스에 대하여 ε-매치하는 데이타 시퀀스를 검색하는데, 이 때 질의 시퀀스로 최소 경계 사각형(MBR: Minimum Bounding Rectangle)을 구성하여 질의 처리를 수행한다. MBR이란 질의 시퀀스의 각 점들의 개수가 너무 많아 개별로 저장할 수 없으므로, 휴리스틱(heuristic)을 사용하여 많은 점들을 포함하는 사각형을 구성하는데, 이 때 최소로 모든 점을 포함하는 사각형을 말한다. 유사 시퀀스 매칭에서는 MBR을 구성한 후, 이 범위 질의(range query)를 통하여 MBR과의 거리가 ε 범위에 있는 데이타 시퀀스의 점을 찾아 후보(candidate) 시퀀스로 정한다. 여기서 후보 시퀀스는 질의 시퀀스와 유사할(ε-매치할) 가능성이 높은 데이타 시퀀스의 서브 시퀀스를 뜻한다.

질의 시퀀스로 MBR을 구성할 때, 하나의 MBR로 구성하는 방법과 여러 개의 MBR로 구성하는 방법으로 나눌 수 있다. 기존의 연구에서는 한 개의 질의 시퀀스를 한 개의 MBR로 구성하여 서브시퀀스 매칭을 하는 방법을 사용하였다. 하지만, 질의 시퀀스의 길이기 긴 경우, 한 개의 MBR만 사용하면 MBR의 크기가 너무 커져서 여러 개의 MBR을 구성하는 것이 효과적이다[4]. 따라서 본 논문에서는 다수의 MBR을 사용하는 서브시퀀스 매칭 방법을 연구하고자 한다. 표 1은 본 논문에서 사용하는 주요 표기와 각 표기에 대한 설명이다.

<표 1> 주요 표기법

기호	설명
Q	사용자에 의해 주어지는 질의 시퀀스
D	데이타 시퀀스
Len(D)	시퀀스 D의 길이

ε	사용자에 의해 주어지는 허용치
ω	슬라이딩/디스조인트 윈도우의 크기
D(Q,S)	시퀀스 Q와 S의 유클리디안 거리 (Q와 S의 길이는 동일함)
p	최소 포함 윈도우 개수[4]

2. 관련 연구

유사 시퀀스 매칭에 대해서는 여러 가지 유사 모델(similarity model)이 연구되었는데, 본 논문에서는 유클리디안 거리에 기반한 모델[1,4]을 사용한다. 길이 n인 두 시퀀스 X = {X[1],X[2],...,X{n]}와 Y = {Y[1],Y[2],...,Y[n]}의 유클리디안 거리 D(X,Y)는 다음과 같이 정의된다.

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(X[i]-Y[i])^2}$$

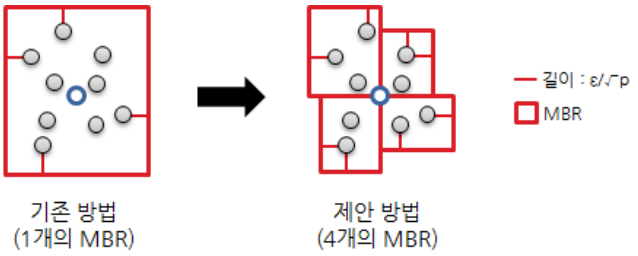
서브시퀀스 매칭은 Faloutsos 등[1]이 제안한 방법과 Dual Match[4]이 가장 대표적인 방법이다. Faloutsos 등[1]이 제안한 방법은 데이타 시퀀스 D를 슬라이딩 윈도우로 나누고, 질의 시퀀스 Q를 디스조인트 윈도우로 나눈 후, MBR을 구성하여 서브시퀀스 매칭을 하는 방법이다. Dual Match는 반대로 데이타 시퀀스 D를 디스조인트 윈도우로 나누고, 질의 시퀀스 Q를 슬라이딩 윈도우로 나눈 후 MBR을 구성하여 서브시퀀스 매칭을 한다. 두 방법 모두 질의 시퀀스로 하나의 MBR을 구성하여 서브시퀀스 매칭을 수행한다.

3. 다수의 MBR을 구성하는 방법

본 논문에서는 서브시퀀스 매칭을 할 때, 질의 시퀀스로 다수의 MBR을 사용하는 방법을 제안 및 구현 하였다. MBR의 개수가 너무 많을 경우 모든 MBR을 매칭 하는데 시간이 더 들기 때문에 오히려 성능이 나빠질 수 있다. 따라서 2~8개의 MBR을 사용하는 것이 적당하다[4]. 따라서 본 논문에서는 질의 시퀀스로 4개의 MBR을 구성하도록 한다. MBR을 구성하기 전에 먼저 질의 시퀀스의 길이가 길기 때문에, Dual Match와 같은 방법으로 DFT 변환[1]을 통해 저차원으로 변환한다. 저차원 변환된 점으로 다수의 MBR을 구성하는 방법으로는 질의 시퀀스의 중심을 기준으로 4개로 나누며, 나눠진 부분에서 $\varepsilon/\sqrt{p}$ 만큼의 공간을 MBR로 구성한다. 여기서 p는 최소포함 윈도우 개수로 시퀀스 D를 크기 ω인 디스조인트 윈도우로 나누었을 때, 길이 Len(D)인 D의 서브시퀀스의 최소포함 윈도우 개수 p는 다음과 같다[4].

$$p = \lfloor (Len(D)+1)/\omega \rfloor - 1 \quad [4]$$

다수의 MBR을 구성할 때, 사용자가 입력한 허용치는 ε이지만 저차원 변환된 점으로 MBR을 구성하므로 각 점보다 $\varepsilon/\sqrt{p}$ 만큼 크게 MBR을 구성해야 한다. 그림 1은 기존의 방법과 제안하는 방법의 MBR 구성 방식을 비교하고 있다. 모든 점보다 $\varepsilon/\sqrt{p}$ 만큼 크게 MBR이 구성되면서도 기존의 MBR보다 작게 구성되는 것을 볼 수 있다.



(그림 1) 다수의 MBR 구성 방식

4. 결과 및 향후 연구

본 논문에서는 시계열 서브시퀀스 매칭 시 질의 시퀀스로 MBR을 구성할 때 다수의 MBR을 구성하는 방식을 제안 및 구현하였다. 다수의 MBR로 구성하여 질의 처리를 하면 질의 시퀀스와 ε-매치 할 가능성이 높은 데이타 시퀀스의 서브시퀀스의 검색 시 개수를 줄일 수 있다. MBR은 저차원 변환 된 값들이 저장되어 있기 때문에 실제 거리를 구하는 비용이 필요한 데, 다수의 MBR을 구성하여 이 비용을 줄일 수 있다. 향후 연구로는 첫째, 다수의 MBR을 구성하는 다양한 방법을 연구하고자 한다. 둘째, 다양한 데이타에 대하여 실험 및 분석을 통해 데이타 특성에 맞는 구성 방법을 찾는 연구를 하고자 한다.

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2011-0002707)

참고문헌

[1] 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

[2] Rafiei D. "On Similarity-Based Queries for Time Series Data" In Proc. the 15th Int'l Conf. on Data Engineering

[3] Rafiei D. and Mendelzon A. "Similarity-Based Queries for Time Series Data" In Proc. Int'l Conf. on Management of Data, ACM SIGMOD

[4] 문양세, 노웅기, 황규영 "윈도우를 구성하는 방법의 이원성을 이용한 효율적인 시계열 서브시퀀스 매칭" 정보과학회논문지:데이타베이스 제28권 제1호, 2001.3, page(s): 1-131