# 시퀀스 데이터베이스를 위한 모양기반의 유사 부분시퀀스 검색

이정화, 윤지희
한림대학교 컴퓨터공학부

# Efficient Retrieval of Similar Shape-Based Subsequences for Sequence Database

Jeong-Hwa Lee, Jee-Hee Yoon
Dept. of Computer Engineering, Hallym University

요약

시퀀스 데이터(sequence data)에서는 각 데이터 값보다는 전후 그 들 사이의 변화 추세 등이 더 큰 정보로 작용하는 것이 일반적이다. 본문에서는 시퀀스 데이터베이스를 대상으로 하여 주어진 시퀀스 패턴과 모양이 유사한 모든 부분시퀀스를 검색해 내는 새로운 방식을 제안한다. 본 방식에서는 시퀀스 데이터의 모양 추출을 위한 데이터 변환, 유사 모양 패턴 클러스터링, 새로운 유사도 계산 방식 등을 도입함으로써, 기존의 방식이 매우 제한적인 패턴만을 유사패턴으로 간주하던 것에 비하여, 패턴이 데이터 축 혹은 타임축으로 각각 확대, 축소, 이동된 경우에도 유사패턴으로 검색이 가능하다.

## 1. 서론

의학, 음악, 과학, 증권 등의 분야에서는 연속 데이터 형태의 시퀀스 데이터(sequence data) 혹은 시계열 데이터(time series database)를 처리 대상으로 하는 컴퓨터 응용 예를 흔히 볼 수 있다. 시퀀스 데이터는 기존의 데이터베이스 응용 분야의 데이터와 다른 특성을 갖고 있어, 시퀀스 상의 각 데이터 값보다는 전후 그 들 사이의 변화 추세 등이 더 큰 정보로 작용할 수 있다. 또한 그 데이터들은 일정한 시간 간격으로 지속적으로 축적되어 일반적으로 그 양이 매우 방대한 것이 특징이다.

시퀀스 데이터베이스에 대한 가장 중요한 질의 처리 유형 중의 하나로, 시퀀스 데이터베이스 상의 유사 시퀀스 검색을 들 수 있다. 즉, 방대한 시퀀스 데이터베이스 내에서 주어진 임의의 시퀀스(패턴) 형태의 질의와 유사한(허용 오차 범위 이내의) 시퀀스를 검색하는 것으로 검색 시퀀스의 출현 유무, 출현 위치 등을 응답 결과로 한다. 이 유사 시퀀스 검색은 응용 목적에 따라 값기반 질의(value-based query)와 모양기반 질의(shape-based query)로 나눌 수 있다. 값기반 질의는 주어진 질의 시퀀스의 데이터 값의 변화 추이 등을 고려한 유사 시퀀스의 검색 방식으로서, "서울의 3월에서 5월간 기온 변화와 유사한 패턴을 갖는 도시를 검색하시오." 등을 예로 들 수 있다. 모양기반 질의는 시퀀스 내의 데이터 값과는 무관하게 주어진 질의 시퀀스와 모양(shape)이 유사한 시퀀스를 검색하는 방식으로서, "24시간 이내의 체온 변화에 급격한 2번의 피크(peak)를 기록하는 Hodgkins 질병의 징후를 갖는 환자를 검색하시오." 등을 예로 들 수 있다.

본문에서는 모양기반의 유사 부분시퀀스 검색기법에 대하여 논한다. 기존의 처리 방식의 한계 등에 대하여 살펴보고, 이를 해결하기 위한 2단계 모양기반 유사 부분시퀀스 기법을 제안한다. 본 방식에서는 처리 대상의 시퀀스 데이터베이스를 모양 추출을 위하여 1차 변환한 후, 주어진 질의 시퀀스 역시 모양 추출 형태로 변환하여, 2차적으로 그들 사이의 유사도 계산에 의하여 허용 오차 범위 이내의 부분시퀀스를 검색해 낸다.

## 2 관련연구

시퀀스 데이터베이스에 대한 유사검색 연구는 주로 값기반 질의처리[1,2,3,4]를 중심으로 이루어지고 있다. [1]에서는 각 시퀀스를 이산 푸리에변환(Discrete Fourier Transform)하여, 다차원 공간상의 한 점으로 매핑, 공간 인덱스 기법 등을 적용함으로써, 공간상의 각 점 사이의 유클리드 거리에 의하여 시퀀스의 유사도를 판별하고 있다. 이 방식은 서로 같은 길이의 시퀀스 사이의 유사도 측정방식인 전체 시퀀스 검색 방식(whole matching method)에 해당하며, [2]에서는 이를 부분 시퀀스 검색 방식(subsequence matching method)으로 확장하고 있다. [3,4]에서는 유사도 측정방식으로 타임와핑(time warping) 기법을 이용하고 있다.

모양기반 질의처리에 관한 연구[5,6,7,8]는 이에 비하여 아직 매우 제한적인 연구 단계이다. [5,6]의 방식에서는 주어

진 시퀀스 데이터베이스를 모양을 보존하도록 정규화 시킨 후 각 시퀀스 사이의 거리측정에 유크리드 혹은 타임와핑 거리를 이용하고 있다. 이 들 방식은 전체 시퀀스 검색방식에 해당한다. [7,8]의 방식에서는 시퀀스데이터베이스의 특징추출을 위하여 이를 함수적 기울기 표현으로 변형하거나 혹은 새로운 모양 정의 언어(Shape Definition Language·SDL)를 이용하여 일파벳 시퀀스로 변형한 후, 인덱싱 기법을 이용하여 원하는 시퀀스를 검색해 낸다 그러나 이 들 방식에서는 임의의 모양을 갖는 부분시퀀스가 데이터축 혹은 타임축으로 각각 확대, 축소, 이동된 경우, 이 들을 검색해 내지 못하는 경우가 자주 발생하며, 또한 응용 의존적이다

## 3 본론

### 3.1 모양기반의 유사 부분시퀀스 검색

문제의 정의는 다음과 같다. 길이가 n인 시퀀스 $S=[S_1,S_2,S_3, \ldots ,S_n]$와 길이가 m인 질의 시퀀스 $Q=[q_1,q_2,q_3, \ldots ,q_m]$가 주어진 경우, S상에서 Q와 모양이 유사한(허용오차 e 이하의) 부분시퀀스 $Si[S_i,S_{i+1},S_{i+2}, \ldots ,S_j]$ $(1\le i<j\le n)$ 를 검색해 낸다. 단, 일반적으로 n>m의 조건이 성립한다. 이때 검색 대상이 되는 유사 부분시퀀스 Si는 Q와 모양이 정확히 일치하는 것은 물론이고, 그 모양이 데이터축 혹은 타임축으로 각각 확대, 축소, 이동된 것을 포함하여야 한다.

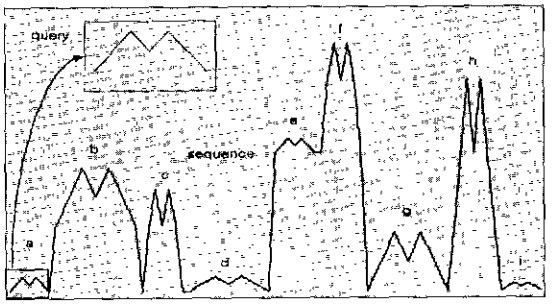


그림 1. 시퀀스 데이터와 질의

즉, 그림 1과 같은 시퀀스 데이터와 질의가 주어진 경우, 유사 부분시퀀스로는 그 모양이 정확히 일치하는 부분시퀀스 a 뿐 아니라 그 모양이 확대(b,c,d,f,g,h), 축소(i), 데이터 축으로 이동(b,e,f), 시간축으로 이동(b,c,d,e,f,g,h,i)된 부분시퀀스 b, c, d, e, f, g, h, i 가 해당된다.

참고문헌[8]의 경우 두 데이터의 차를 미리 정의된 알파벳의 심볼로 변형하여 새로운 알파벳 시퀀스를 생성하므로, 값의 변화가 큰 c, f, h (값축으로 확대)는 질의와 다른 알파벳 시퀀스로 변형되어 질의 응답 결과로 얻어지지 못한다. 또한 참고문헌[7]의 경우 일련의 부분시퀀스를 함수의 기울기로 표현하므로 값의 변화가 작은 축소된 부분시퀀스인 i에 대한 정보가 손실되어, 원하는 결과 값을 얻지 못하는 경우가 발생한다.

이들 문제점을 해결하기 위하여 본 논문에서는 효율적인 유사 검색을 위한 2 단계의 모양기반 질의 처리 방법을 제시한다 제시된 방법의 첫 단계는 숫자타입의 시퀀스 데이터를 모양을 보존하도록 분류하여 새로운 시퀀스 데이터를 생성 하며, 두 번째 단계에서는 새롭게 생성된 시퀀스 데이터와 주어진 질의와의 유사도를 측정하여 원하는 결과들 검색한다

### 3.2 모양 분류 및 클러스터된 데이터 생성

주어진 시퀀스 데이터는 전반적인 모양을 알 수 없는 숫자타입으로 되어있으므로 모양을 표현할 수 있는 새로운 타입의 데이터로 전환이 필요하다

길이가 n인 시퀀스 데이터 $S=[S_1, S_2, .., S_i, .. S_n]$가 주어졌을 때 $S_i$ $(i=(k+1)/2, \ldots ,n-(k-1)/2)$를 중심으로 길이가 k인 데이터를 하나의 모양으로 간주하여 $n-k+1$ 개의 모양 표현 데이터 $R=[R_1, R_2, .., R_j, .., R_{n-k+1}]$로 변환 한 후 모양별로 유사한 데이터를 하나의 클래스로 간주하는 클러스터링을 실행, 최종 변환 데이터를 생성한다.

모양 표현 데이터 $R_j=[r_{j1}, .. r_{jk-1}]$는 두 데이터의 차에 대한 부호로 표현되므로 모두 세 개의 심볼( +, -, 0 )을 가질 수 있으며, 서로 다른 모양 표현 데이터 $R_j$ 는 최대 $3^{k-1}$개이다 즉, k가 3일 경우 모양 표현 데이터 $R_j$ 는 다음과 같다.

$$S = < .., S_{i-1}, S, S_{i+1}, S_{i+2}, \ldots >$$

$$R_j = < \mathrm{sign}(S_{i-1} - S_i), \mathrm{sign}(S_{i+1} - S_i) >$$

기존에 연구되어져 있는 클러스터링(clustering) 방법은 k-평균(k-means), 최대-최소 거리(min-max distance), 분할-합병(ISODATA) 등 많은 알고리즘이 있으나, 이는 모두 데이터의 빈도수와 값에 의존하여 클러스터링을 하므로 [9] 모양 분류에 따른 시퀀스 데이터에는 적합하지 않다

그러므로, 유사한 모양을 하나의 클래스로 간주하는 새로운 클러스터링 방법이 필요하며, 본 논문에서는 새롭게 생성된 시퀀스 데이터에 적합하도록 표 1과 같이 클래스를 정의하여 클러스터링을 실행하였다.

표 1 모양 표현 데이터에 따른 클래스 분류

| sign(si-1 - si) | sign(si+1 - si) | class |
|---|---|---|
| + | + | a |
| + | - | b |
| - | + | c |
| - | - | d |

예를 들어 시퀀스 데이터 S=<1.31, 2.64, 3.19, 2.32, 3.20, 2.87, 1.06>와 모양 표현 데이터 R의 길이 k가 3으로 주어졌을 경우 표 1을 참조하여 클러스터링을 실행한 후 생성된 새로운 시퀀스 데이터 C 는 다음과 같다.

S=<1.31, 2.64, 3.19, 2.32, 3.20, 2.87, 1.06>
$R_1$=<-,+> $R_2$=<-,-> $R_3$=<+,+> $R_4$=<-,-> $R_5$=<+,->
C=< c, d, a, d, b >

### 3.3 유사도 측정 및 검색

숫자 타입으로 구성된 두 시퀀스 데이터의 유사도를 측정하는 가장 일반적인 방법은 유클리디안 거리(Euclidian distance)를 사용하는 방법[1][2]이며, 데이터의 확대나 축소 상황을 고려한 타임와핑(time-warping) 거리측정이[3][4] 연구되어져 있으나, 클러스터링을 실행한 후 클래스타입의 데이터로 새롭게 구성된 시퀀스 데이터의 유사도를 측정하기에는 적합하지 못하다.

또한 문자타입의 두 패턴(pattern)의 유사도를 측정하는 일반적인 방법은 유전자 알고리즘에서 주로 사용되는 에디트 거리(edit distance)측정방법[10] 있으나, 패턴의 길이를 고려하여 거리를 측정하므로 서로 다른 길이의 확장된 데이터를 검색해야 하는 본 논문의 목적에는 적합하지 않다.

타임와핑 거리측정 방법과 에디트 거리측정 방법은 모두 동적 계획법을 이용하여 두 패턴 요소간의 대응을 수행하여 유사도를 계산하는 DP(Dynamic Programming)를 기반으로 하였으며[11], 본 논문에서 제시하고자 하는 거리 측정 방법 또한 패턴의 길이에 제한을 받지 않는 DP를 기반으로 하여 확대, 축소, 이동된 모든 부분시퀀스와의 유사도 측정시 적절한 거리 값을 가지도록 한다

즉, 유사도를 측정하기 위한 두개의 시퀀스 데이터 $C1=[c_1, c_2, .., c_n]$, $C2=[c_1, c_2, ..., c_m]$가 주어졌을 경우 $C_i$ 와 $C_j$의 거리 $D(i,j)$는 다음과 같다.

$$D(i,j) = \min[D(i-1,j) + 1, D(i, j-1) + 1, D(i-1, j-1) + t(i,j)]$$

$i=0$ 이거나 $j=0$일 경우 $D(i,j)$는 무한대($\infty$)이며, $t(i,j)$는 $C1(i)$와 $C2(j)$가 같을 경우 0, 그렇지 않을 경우 1을 갖는다.

두 시퀀스 데이터의 유사도는 거리가 작을수록 커지므로, 질의의 최종 결과는 질의와의 거리가 주어진 허용오차 이내인 모든 부분시퀀스이다.

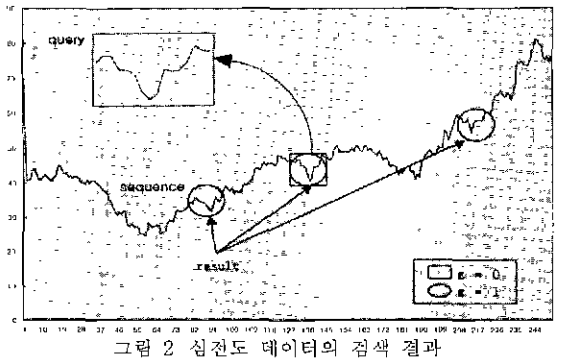


그림 2 심전도 데이터의 검색 결과

## 4. 실험

본 논문에서의 실험 데이터는 확대, 축소, 이동의 모든 상황을 포함하도록 임의로 구성한 그림 1의 시퀀스 데이터와 그림 2의 실제 심전도 데이터를 사용하였다.

첫 번째 실험은 기존에 제시된 모양기반 질의 처리 방법의 하나인 SDL[8]과 본 논문에서 제시한 질의 처리 방법의 비교 실험이다. SDL의 알파벳 A는 표 2와 같이 정의되며, 본 논문에서 제시한 방식의 모양 표현 데이터 길이 k는 3, 허용 오차는 0으로 주어졌다.

SDL의 경우 값으로만 확대된 c, f, h 즉, 두 데이터의 차가 큰 경우 서로 다른 심볼을 가지므로 질의 응답 결과를 얻을 수 없었으나, 본 논문에서 제시한 방식의 경우에는 (k=3) 유사한 모든 부분시퀀스를 질의 응답으로 얻을 수 있었다.

두 번째 실험은 표현 데이터의 길이 k를 3 과 5로 주었을 경우의 비교 실험이며, 허용오차는 모두 0 으로 주어졌다. k가 3일 경우 a, b, c, d, e, f, g, h, i 의 원하는 모든 부분시퀀스를 찾았으나, k가 5일 경우 b, d, g와 같이 타임축으로 확대된 경우 찾지 못했다.

이 실험 결과를 살펴보면, 그림 1의 시퀀스 데이터의 경우에는 효율적인 질의 응답을 얻기 위해서는 k의 값이 3일때 가장 적합하다는 것을 알 수 있다.

즉, 길이가 k인 모양 표현 데이터 $R_i$를 얻기 위한 k값은 응용 분야와 데이터에 의존적이며, 가장 효율적인 결과를 얻기 위한 일반적인 k를 구하는 알고리즘 개발이 필요하다

마지막으로 실제 심전도 데이터를 사용하여 본 논문에서 제시한 방법으로 유사도 거리 측정 시 허용 오차($\varepsilon$)를 각각 0과 1로 주어 실험하였으며 결과는 그림 2에 나타난 것처럼 시퀀스 내의 모든 부분시퀀스와 유사도를 측정하여 주어진 질의와 유사한 모양의 모든 부분시퀀스를 얻을 수 있었다.

표 2. 알파벳 A

| 심볼 | 설명 | 데이터의 차이 | |
|---|---|---|---|
| | | 하위값 | 상위값 |
| Up | 급격하게 상승 | 5.01 | 20.0 |
| up | 완만하게 상승 | 0.01 | 5.0 |
| zero | 변화 없음 | 0 | 0 |
| down | 완만하게 하강 | -0.01 | -5.0 |
| Down | 급격하게 하강 | -5.01 | -20.0 |

## 5 결론 및 향후 연구 계획

유사 시퀀스 검색은 응용목적에 따라 값기반 질의와 모양 기반 질의로 나누어 질 수 있다. 본 논문에서는 기존에 연구되어졌던 모양 기반 질의 처리의 문제점을 지적하고, 해결하는 방안으로서 유사 시퀀스 검색의 가장 중심적인 문제인 확대, 축소, 이동의 상황을 부분시퀀스 검색에서 고려한 2 단계의 새로운 모양기반 질의 처리 방법을 제시하였다.

제시된 방법에서는 처리 대상의 시퀀스 데이터를 모양 추출을 위하여 1차 변환한 후, 주어진 질의 시퀀스 역시 모양 추출 형태로 변환하여, 2차 적으로 그들 사이의 유사도 계산에 의하여 허용오차 범위 이내의 부분시퀀스를 검색하였다.

새로운 질의 처리 기법은 기존에 연구되어졌던 모양기반 질의에서 해결하지 못한 문제점을 해결하였으나, k 값에 독립적으로 유사한 모양을 하나의 클래스로 간주하는 클러스터링 알고리즘의 구현에 관한 연구가 새로운 과제로 남아있다.

또한, 시퀀스 데이터베이스는 그 양이 매우 방대하므로 검색의 효율을 높이기 위하여 새롭게 생성된 시퀀스 데이터에 적합한 인덱스 구성에 관한 연구를 수행할 계획이다.

## 참고문헌


[1] R Agrawal, C Faloutsos, and A. Swami, "Efficient Similarity Search in Sequence Database", In Proc, Intl, Conf, on Foundations of Data Organization and Algorithms, 1993

[2] C Faloutsos, M Ranganathan, and Y Manolopoulos, "Fast subsequence matching in time-series database", In Proc, Intl, Conf, ACM SIGMOD, May 1994.

[3] Byoung-kee Yi H V. Jagadish, and C Faloutsos, "Efficient retrieval of similar time sequences under time warping", In Proc, Intl, Conf, on Data Engineering, 1998

[4] T Bozkaya, N. Yazdani and M Ozsoyoğlu. "Matching and indexing sequences of different lengths", In Proc, CIKM, 1997.

[5] D Q Goldin and P. C. Kanellakis, "On similarity queries for time-series data constraint specification and implementation", In Proc, of Constraint Programming, September 1995

[6] Davood Rafiei and Alberto Mendelzon, "Similarity- based queries *for time series data*", *In Proc, Conf, ACM SIGMOD*, 1997.

[7] H Shatkay and S B. Zdonik,"Approximate queries and representation for large data sequences", In Proc, Conf, of Data Engineering, February 1994.

[8] R Agrawal, G Psaila, E L. Wimmers and M Zait, "Querying shapes of histories", In Proc, Conf, VLDB, 1995

[9] Julius T. Tou, and Rafael C Gonzalez, "Pattern Recognition Principles", Addison-Wesley Pub pp.75-109,1974

[10] Dan Busfield, "Algorithms on strings, trees and sequences computer science and computational biology" Cambridge University Pr, pp 215-253, 1997

[11] 김삼운,"식별 알고리즘을 중심으로 한 패턴인식 입문" 홍릉 과학 출판사, pp 119-136,1995