시계열 데이터베이스에서 순위를 지원하는 서브시퀀스 매칭 방법을 위한 시각화 툴≠

이성진*, 이진수*, 조훈‡, 한욱신*
*경북대학교 컴퓨터공학과
‡경북대학교 의료정보원천기술연구소
sjlee@www-db.knu.ac.kr, jslee@www-db.knu.ac.kr,
hunecho@knu.ac.kr, wshan@knu.ac.kr

A Visualization Tool for Ranked Subsequence Matching in Time-Series Databases≠

Sung-Jin Lee*, Jinsoo Lee*, Hune Cho‡, Wook-Shin Han*
*Department of Computer Engineering, Kyungpook National University
‡Medical Informatics Platform for Telehealth Development Center,
Kyungpook National University

요 약

시계열 데이터(time-series data)는 연속적인 데이터를 고정된 시간 간격으로 샘플링한 실수 값들의 연속을 의미한다. 시계열 데이터의 예로는, 음악 및 동영상 데이터, 심전도 데이터, 주식 그래프 등의 데이터가 있다. 시계열 데이터는 다시 데이터베이스에 저장 되어있는 데이터 시퀀스(data sequence)와, 사용자에 의해 주어지는 질의 시퀀스(query sequence)로 분류된다. 시계열 데이터베이스(time-series database)에서 순위를 지원하는 서브시퀀스 매칭 방법(ranked subsequence matching)은 데이터 시퀀스와 질의 시퀀스가 주어졌을 때, 질의 시퀀스의 길이와 같은 데이터 시퀀스의 서브시퀀스(subsequence)들 중에서 질의 시퀀스와 가장 유사한 상위 k개의 서브시퀀스들을 찾는 것이다. 본 논문의 목적은 사용자가 매칭 방법에 대한 인식과 이해가 부족하더라도 기존의 콘솔 기반의 매칭 프로그램을 보다 쉽게 사용할 수 있도록 이용성을 향상시키기 위하여 시각화 툴을 개발하는 것이다. 구체적으로, 5가지 시각화(visualization) 기능을 제공하는 사용자 인터페이스를 구현하였다. 구현된 사용자 인터페이스를 통해 사용자가 기존의 매칭 프로그램을 보다 쉽고 간편하게 사용할 수 있도록 기여한다.

1. 서론

최근 시계열 데이터에서 순위를 지원하는 서브시퀀스 매칭 방법[1]이 연구되었다. 관련된 연구로 [2,3,4,5,6,7,8,9] 등이 있다. 그러나 일반 사용자가 이러한 매칭 방법을 이용한 프로그램을 자신이 속한 분야에 적용하여 사용하기에는 매칭 방법의 이해와 프로그램 사용을 위한 다소의 전문성이 요구되고 그에 따른 어려움이 존재해 왔다. 구체적인 예로, 사용자가 콘솔 기반의 매칭 프로그램을 사용할 경우, 콘솔 창을 띄우고 프로그램 실행을 위한 명령어를 손으로 직접 입력하는 과정을 거쳐야 하고, 그에 따른 결과를 다시 분석하여 보기 위해서는 실제 결과 값을 나열해서 보거나 그래프 형태로 변환하여 보는 수밖에 없는 어려움이 있다. 이러한 기존의 매칭 프로그램을 사용하기 위한 이용성의 문제점을 해결한다면 더욱 많은 분야에서 매칭 프로그램을 사용할 수 있을 것이다. 하지만, 아직 기존의 매칭 방법을 이용한 프로그램 사용의 이용성의 단점을 극복하기 위한 연구는 활발하게 이루어지지 않았다.

본 논문에서는 기존 매칭 프로그램의 이용성의 단점을 극복하기 위해 시각화 방법을 제안하고, 이를 위한 사용자 인터페이스를 구현하였다. 또한, 사용자가 시계열 데이터 입력을 통한 질의를 하고 질의에 대한 결과를 볼 때, 최대한 이해하기 용이하도록 시각화된 데이터의 확대, 축소, 이동 기능 등 부가적인 기능을 추가하였다.

≠ 본 논문은 지식경제부 산업원천기술개발사업의 “건강서비스 적정 보상체계 및 온톨로지 DB 모델링 기술“ 에서 지원된 연구임.

본 논문의 나머지는 다음과 같이 구성된다. 제 2장에서는 본 논문에서 이용하는 관련 데이터 및 알고리즘에 대해 소개하고, 제 3장에서는 본 논문에서 제안한 시각화 프로그램 구현 내용에 대해 설명한다. 마지막으로, 제 4장에서는 본 논문의 결론에 대해 설명한다.

2. 관련연구

제 2장에서는 기존의 연구 중 본 논문과 관련된 시계열 데이터 및 알고리즘들에 대해 간략히 소개한다.

2.1. 시계열 데이터(Time-Series Data)

시계열 데이터[2,3,4,5,6,7]는 연속적인 데이터를 고정된 시간 간격으로 샘플링한 실수 값들의 연속을 의미한다. 데이터 시퀀스는 데이터베이스에 저장 되어있는 시계열 데이터를 의미하고, 질의 시퀀스는 데이터 서브 시퀀스 매칭을 위하여 사용자에 의해 주어지는 시계열 데이터를 의미한다.

2.2. DTW(Dynamic Time Warping)

DTW[8,9]는 두 시퀀스 간의 유사도를 측정하는 유사도 검사 알고리즘이다. 매칭 프로그램에서 데이터 시퀀스와 질의 시퀀스가 모두 입력되었을 때, 질의 시퀀스와 데이터 시퀀스의 서브시퀀스가 서로 얼마나 유사한지 측정하기 위해 사용한다. 측정된 유사도 값은 질의 시퀀스에 대한 서브시퀀스들의 유사도 순위를 계산하기 위해 사용된다.

2.3. 순위를 지원하는 서브시퀀스 매칭

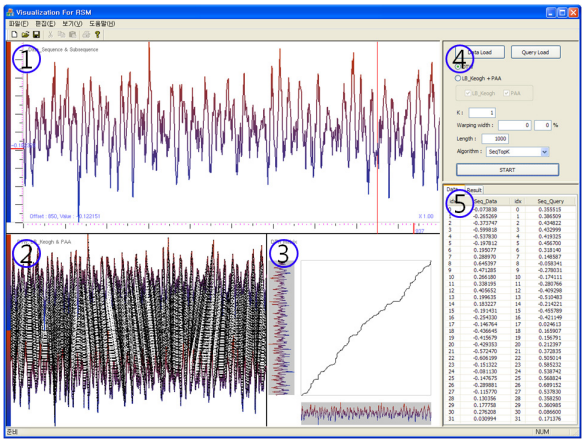
순위를 지원하는 서브시퀀스 매칭 알고리즘[1]은 데이터

시퀀스와 질의 시퀀스가 주어졌을 때, 질의 시퀀스의 길이와 같은 데이터 시퀀스의 서브 시퀀스들 중에서 질의 시퀀스와 가장 유사한 상위 k개의 서브시퀀스들을 찾는 것이다. 본 논문에서는 [1]에서 제안한 SeqTopK, AdvTopK, RangeTopK, DualMatchTopK 그리고 DeferredTopK 다섯 가지의 매칭 알고리즘을 사용한다. 각 알고리즘에 대한 자세한 설명은 공간의 제약으로 인해 본 논문에서 생략한다.

3. 사용자 인터페이스 구현

제 3장에서는 앞에서 언급한 기존 연구방법들을 기반으로, 실제 데이터 값들과 매칭된 결과값들을 이용하는 시각화 프로그램을 구현한 내용을 설명한다.

사용자 인터페이스(Graphic User Interface, GUI)는 질의 및 데이터 시퀀스를 입력받고 [1]에서 제안된 알고리즘을 통해 매칭 프로그램을 수행한다. 이후 입력받은 데이터를 시각화하여 화면에 보이도록 한다.



[그림 1] 사용자 인터페이스 화면

위 [그림 1]은 사용자 인터페이스를 통해 실제 데이터를 입력 하였을 때의 예시 화면이다. 각 화면과 기능에 대한 설명은 아래 <표 1>에서 간략히 정리하였고, 자세한 내용은 <표 1> 아래 다시 설명한다.

<표 1> 사용자 인터페이스 화면에 대한 설명

위치	화면	설명
①	데이터 시퀀스 화면	데이터 시퀀스 값을 화면에 그려줌 확대/축소/이동 가능
②	시퀀스 대조 화면	DTW 알고리즘을 이용해 두 시퀀스의 매칭되는 값을 화면에 그려줌
③	DTW 매트릭스 화면	DTW 매트릭스 화면을 그려줌
④	옵션 화면	데이터 입력 및 프로그램 실행
⑤	데이터 탭 화면	실제 입력 값과 결과 값을 출력

데이터 시퀀스 화면은 데이터 시퀀스를 입력했을 때, 화면상에 입력된 데이터 시퀀스를 그려준다. 화면에서 보여지는 시퀀스는 실제 저장된 값의 크기에 따라 위치와 색이 달라진다. 값이 클수록 위쪽에 위치하고 붉은색에 가까워지며, 값이 작을수록 아래쪽에 위치하고 파란색에 가까워진다. 확대/축소/이동 기능을 통하여 데이터 시퀀스의 값을 더욱 세밀히 확인할 수 있게 하였고, 마우스 움직임에 따라 마우스 포인터가 가르키는 점의 위치에 해당하는 실제값(value)과 실제값의 오프셋(offset)을 실시간으로 화면상에 보이도록 하였다. 해당 점을 마우스 왼쪽 클릭 했을 경우, 고정된 붉은색 세로 실선이 그려지고 실선에 접하는 데이터 시퀀스의 값과 오프셋 값을 화면상에서 사용자에게 보여준다.

시퀀스 대조 화면은 데이터 시퀀스와 질의 시퀀스가 모두 입력되었을 때, DTW 알고리즘을 이용하여 서로 매칭되는 값들을 선으로 이어 사용자가 시각적으로 확인 가능하도록 하였다.

DTW 매트릭스 화면은 화면에서 보이는 사각형 좌측 아래를 시작으로 우측 위 끝으로 이어지는 대각선 방향의 중심에 선이 가까울수록 두 데이터의 값이 유사함을 의미한다.

옵션 화면은 사용자가 질의하고자 하는 데이터를 입력하고 프로그램을 실행하여 그에 대한 결과를 화면상에 보이도록 해준다.

데이터 탭 화면은 데이터 탭과 결과 탭으로 구성되어있다. 입력된 실제 데이터 시퀀스 값과 질의 시퀀스 값의 리스트를 데이터 탭에서 모두 보여주고 매칭된 서브 시퀀스 데이터의 순위 및 위치 값을 결과 탭에서 사용자에게 보여준다.

4. 결론

본 논문에서는 현재까지 연구된 시계열 데이터베이스에서 순위를 지원하는 서브시퀀스 매칭방법을 사용자 인터페이스를 통해 시각화하여 사용자에게 보여줌으로써 기존 매칭 프로그램의 이용성의 단점을 극복하고 사용자가 이전 연구들에서 구현된 매칭 알고리즘을 모두 이해하지 못하더라도 쉽고 편리하게 시계열 데이터들을 비교 분석할 수 있도록 기여하였다.

참고문헌

[1] Han, W., Lee, J., Moon, Y., and Jiang H. “Ranked Subsequence Matching in Time-Series Database,” In VLDB, pp. 423-434, 2007.

[2] Agrawal, R., Faloutsos, C., and Swami, A., “Efficient Similarity Search in Sequence Databases,” In FODO, 1993.

[3] Faloutsos, C., Ranganathan, M., and Manolopoulos, Y., “Fast Subsequence Matching in Time-Series Databases,” In SIGMOD, pp. 419-429, 1994.

[4] Moon, Y., Whang, K., and Loh, W., “Duality-Based Subsequence Matching in Time-Series Databases,” In ICDE, pp. 263-272, 2001.

[5] Rafiei, D. et al., “Querying Time Series Data Based on Similarity,” IEEE TKDE, Vol. 12, No. 5, 2000.

[6] Lim, S.-H. et al., “Using Multiple Indexes for Efficient Subsequence Matching in Time-Series Databases,” In DASFAA, pp. 65-79, 2006.

[7] Keogh, E. “A Decade of Progress in Indexing and Mining Large Time Series Databases,” In VLDB, Tutorial, 2006.

[8] Berndt, D. and Clifford, J., “Finding Patterns in Time Series: a Dynamic Programming Approach,” In Advances in Knowledge Discovery and Data Mining, pp. 229-248, AAAI/MIT, 1996.

[9] Sakoe, H. and Chiba, S, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” IEEE Trans. on ASSP, Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.