다중윈도우를 이용한 서브시퀀스 매칭 방법 구현

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

An Implementation of a Subsequence Matching Method for Multiple Windows

Ah-Yeon Jin*, Young-Ho Park*
*Dept of Multimedia Science, Sookmyung Women's University

요 약

시계열 데이터는 기상데이터, 주식데이터, 센서 데이터, 네트워크 트래픽 데이터, 의료 데이터 등 다양한 분야에서 사용되고 있다. 그 중에서 서브시퀀스 매칭 방법은 시계열 데이터베이스 어플리케이션에서 많은 주목을 받고 있다. 기존의 서브시퀀스 매칭 방법은 단일 윈도우만을 비교하여 서브시퀀스 매칭을 수행하였으나, 착오해답을 줄이는 데에는 한계가 있었다. 따라서 다중 윈도우를 비교하여 착오해답을 줄이고 성능을 높일 수 있는 다중 윈도우를 이용한 서브시퀀스 매칭 방법을 구현하였다. 그 결과 단일 윈도우를 사용했을 때보다 약 4.8 배까지 후보집합의 수가 줄어드는 것을 볼 수 있었다.

1. 서론

시계열 데이터는 특정 시간에 측정된 실수 값들의 시퀀스를 의미한다.[2,3,4]. 시계열 데이터는 기상데이터, 주식데이터, 센서 데이터, 네트워크 트래픽 데이터, 의료 데이터 등 다양한 분야에서 사용되고 있다. 이처럼 다양한 분야에서 사용되는 시계열 데이터베이스 어플리케이션에서 가장 중요한 부분은 서브시퀀스 매칭 방법이다. 서브시퀀스 매칭 방법은 지속적으로 연구되고 있지만 주로 단일 윈도우를 이용한 방법이 연구되고 있다.

본 논문에서는 서브시퀀스 매칭 방법에 대한 관련 연구를 실시하고, 기존의 단일 윈도우만을 이용한 서브시퀀스 매칭방법에서 발전한 다중 윈도우 서브시퀀스 매칭 방법을 구현한다.

2. 관련연구

가장 잘 알려진 서브시퀀스 매칭 방법은 Faloutsos 등(이하 FRM)[1]과 DualMatch[2]이다. FRM 은 데이터 시퀀스를 슬라이딩 윈도우로 분할하고 질의 시퀀스를 디스조인트 윈도우로 분할하여, 그 윈도우들을 f 차원 점으로 변환시킨다. 슬라이딩 윈도우의 저차원 변환점은 MBR(minimum bounding rectangle)로 변환하여 인덱스를 구성하고, 디스조인트 윈도우의 저차원 변환점은 각 개별 점에 대해 $\varepsilon/\sqrt{p}$ 만큼의 범위 질의를 수행하여 후보집합을 구하는 방법이다.

인덱스 구성 단계에서 착오해답이 많이 발생하는 문제점을 해결하기 위하여 DualMatch 가 제안되었다. DualMatch 는 윈도우 구성 방법에 대해 FRM 의 이원적 접근법에 해당한다. FRM 과 달리 데이터 시퀀스를 디스조인트로 나누고 질의 시퀀스를 슬라이딩 윈도우로 분할하여, 그 윈도우들을 f 차원으로 저차원 변환시킨다. 데이터 시퀀스의 윈도우의 저차원 변환점을 MBR 로 구성하는 경우 착오해답이 많이 발생하였기 때문에, 슬라이딩 방법보다 윈도우 개수가 적은 디스조인트 윈도우로 나누고 MBR 로 변환하는 대신 개별점으로 인덱스를 구성한다. 그러나 질의 시퀀스의 윈도우의 저차원 변환점은 MBR 로 변환해 질의 시퀀스를 수행하도록 하여, 착오해답을 줄이고 최종 후보집합을 구하는 방법이다.

3. 다중 윈도우를 이용한 서브시퀀스 매칭방법

서브시퀀스 매칭에서 착오해답은 서브시퀀스 매칭의 후처리 단계에서 많은 디스크 액세스와 CPU 작업량을 차지한다. 따라서 착오해답을 줄이는 것은 서브시퀀스 매칭 방법의 효율성을 높일 수 있는 가장 좋은 방법이다. DualMatch 는 윈도우 구성 방법에 대해 이원적으로 접근함에 따라 FRM 의 착오해답을 획기적으로 줄였으나 여전히 많은 착오해답을 발생시키고 있다.

이를 해결하기 위한 방안 중의 하나로 단일 윈도우만을 사용하던 기존과 달리 다중 윈도우를 비교함으로써 착오해답을 줄이는 방안을 구현하였다.

단일 윈도우를 사용한 DualMatch 에서는 질의 시퀀스를 슬라이딩 윈도우로 분할하여 저차원 변환후 MBR 을 구성한다. 그 후 각 윈도우의 변환된 MBR 을 이용하여 범위 질의를 수행하여 바로 최종 후보집합을 구하도록 되어있지만, 다중 윈도우를 이용하기 위해서는 범위 질의 수행 후 한가지 단계를 더 추가하였다. 1 차적으로 MBR 을 이용한 범위질의를 수행하여 1 차 후보집합을 가진다. 2 차적으로, 후보집합을 기반으로 해당 윈도우와 매치된 부분을 제외한 나머지 질의 시퀀스부분에 대하여 필터링 과정을 거친다. 즉, 질의 시퀀스 나머지 부분의 각 점과 이와 매치하는 데이터 시퀀스 윈도우의 각 점의 거리를 계산하여, 한계값 이내에 속한다면 최종적인 후보집합으로 할당시킨다.

4. 성능평가

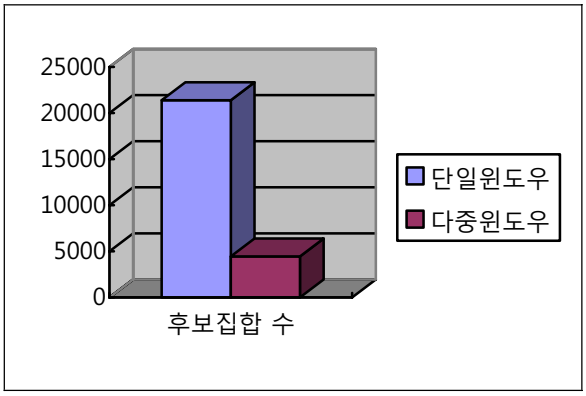
본 절에서는 DualMatch 에서 단일 윈도우를 사용했을 때와 다중윈도우를 사용했을 때를 비교한 성능평가 결과를 설명한다.

4.1 실험환경 및 데이터

본 실험에서는 C 언어로 단일 윈도우를 사용한 DualMatch 와 다중윈도우를 이용하는 서브시퀀스 매칭방법을 구현하였고, 인텔 i5-750 쿼드 코어 프로세서를 사용하고 16GB 의 메인메모리를 가진 2.80GHz 리눅스 PC 에서 실험을 수행하였다. 또한 1,055,525 엔트리로 구성된 합성 데이터셋 MIX-DATA[1]를 이용하여 비교 실험하였다.

4.2 실험결과

본 실험에서는 하나의 질의 시퀀스를 수행하여 후보집합의 수를 통해 단일 윈도우를 사용했을 때와 비교하였다. 비교결과 약 4.8 배까지 후보집합의 수가 줄어든 것을 볼 수 있다.



(그림 1) 비교윈도우 수에 따른 후보집합 수

5. 결론

본 논문에서는 단일 윈도우만을 비교하던 기존의 서브시퀀스 매칭 방법과 달리 다중 윈도우를 비교함으로써 착오해답을 줄이는 방안에 대하여 구현하였다. 구현한 내용을 바탕으로 단일 윈도우를 사용했을 때와 다중윈도우를 사용했을 때를 비교하여 성능평가를 한 결과, 약 4.8 배까지 후보집합의 수가 줄어드는 것을 볼 수 있었다.

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

참고문헌

[1] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, Fast subsequence matching in time-series databases. Proceedings of the 1994 ACM SIGMOD international conference on Management of data, 1994.

[2] Yang-Sae Moon, Kyu-Young Whang, and Woong-Kee Loh, Duality-based subsequence matching in time-series databases. Proceedings of the 17th International Conference on Data Engineering, 2001.

[3] Yang-Sae Moon, Kyu-Young Whang, and Woong-Kee Loh, General match: a subsequence matching method in time-series databases based on generalized windows. Proceedings of the 2002 ACM SIGMOD international conference on Management of data, 2002.

[4] Chotirat Ann Ratanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos and Gautam Das, Mining time series data. Data Mining and Knowledge Discovery Handbook, 2010.