시계열 데이터베이스에서 유사 시퀀스 매칭 방법에 관한 조사

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

A Survey on Similar Sequence Matching Methods in Time-Series Database

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

요 약

시계열 데이터는 경제, 기상, 의료 등 다양한 분야에서 사용되고 있으며, 시계열 데이터 상에서의 검색 방법에 대한 관심이 더욱 높아지고 있다. 시계열 데이터는 각 시간별로 측정한 실수 값의 시퀀스로, 사용자가 원하는 질의 시퀀스와 유사한 데이터 시퀀스를 찾는 방법인 유사 시퀀스 매칭 방법을 조사한다. 유사 시퀀스 매칭 방법은 전체 매칭과 서브시퀀스 매칭으로 분류되며, 서브시퀀스 매칭의 대표적인 방법으로 전체매칭을 일반화한 방법인 FRM, FRM 의 윈도우 구성 방법에 대해 이원적으로 접근한 DualMatch, FRM 과 DualMatch 를 일반화한 GeneralMatch 가 있으며, 각 방법에 대한 비교분석을 한다.

1. 서론

시계열 데이터는 주식, 환율변동 같은 경제분야, 온도, 강수량변화 같은 기상분야뿐만 아니라 심전도, 뇌전도와 같은 의료분야 등 다양한 분야에서 사용되고 있다.[1,2,3,5] 이처럼 많은 분야에서, 주어진 데이터를 응용하기 위해 시계열 데이터 검색에 대한 관심이 더욱 높아지고 있다.

본 논문에서는 시계열 데이터와 유사 시퀀스 매칭에 대한 정의를 내린다. 또한 유사 시퀀스 매칭 방법을 전체 매칭과 서브시퀀스 매칭으로 나누어 설명하고, 전체 매칭의 방법(Agrawal 등[2])과 서브시퀀스 매칭의 대표적인 방법인 Faloutsos 등(이하 FRM)[1], DualMatch[3], General Match[4]에 대해 조사하고, 각 방법에 대한 비교분석을 한다.

2. 시계열 데이터

시계열 데이터는 각 시간별로 측정한 실수 값의 시퀀스로, 데이터베이스에 저장된 시계열 데이터를 데이터 시퀀스(data sequence)라 한다. 이러한 데이터 시퀀스 상에서 사용자가 원하는 질의 시퀀스(query sequence)와 유사한 데이터 시퀀스를 찾는 방법을 유사 시퀀스 매칭(similar sequence matching)이라고 한다.[3] 다음 장에서는 유사 시퀀스 매칭 방법을 전체 매칭과 서브시퀀스 매칭으로 분류하여 대표적인 방법들과 함께 자세히 설명한다.

3. 유사 시퀀스 매칭 방법

본 절에서는 유사 시퀀스 매칭 방법을 전체 매칭과 서브시퀀스 매칭으로 나누어 대표적인 방법을 설명한다.

3.1 전체매칭

유사 시퀀스 매칭 방법은 Agrawal 등[2]에 의해 처음 소개되었으며, 이는 데이터 시퀀스와 질의 시퀀스의 길이가 동일한 경우 사용되는 방법인 전체 매칭 방법을 사용하였다. 데이터 시퀀스를 이산 퓨리에 변환(DFT: discrete fourier transform)하여 f 개의 특성을 추출하고, 이를 f 차원의 R*-트리에 저장한다. 질의 시퀀스 또한 f 차원의 점으로 변환하여 허용치 ε 만큼의 범위 질의를 구성한다. 해당 범위 질의로 R*-트리를 검색하여 후보집합을 구하는 방법이다.

3.2 서브시퀀스 매칭

본 절에서는 서브시퀀스 매칭의 대표적인 방법인 FRM, DualMatch, GeneralMatch 에 대해서 설명한다.

3.2.1 FRM

FRM 은 전체 매칭을 일반화한 방법으로, 인덱스 구성단계와 서브시퀀스 매칭 단계로 이루어져있다. 인덱스 구성단계는 데이터 시퀀스를 슬라이딩 윈도우로 나누고, 그 슬라이딩 윈도우를 f 차원으로 변환한 값을 R*-트리에 MBR(minimum bounding rectangle)로 저장하는 단계이다. 서브시퀀스 매칭단계는 질의 시

퀀스를 p(=⌊(Len(S))/ω⌋)개의 디스조인트 윈도우로 나누어 f 차원 점으로 변환하고, 개별 점에 대해 $\varepsilon/\sqrt{p}$ 만큼의 범위 질의를 구성한다. 해당 범위 질의를 이미 구성된 인덱스에서 검색하여 후보집합을 구하는 방법이다.

인덱스 구성단계에서 개별 점대신 MBR 을 저장하기 때문에 착오해답(false alarms)이 많이 발생하는 문제점이 생긴다. 이러한 문제점을 해결하기 위한 방법으로 DualMatch 가 제안되었다.

3.2.2 DualMatch

DualMatch 는 윈도우 구성 방법에 대하여 FRM 의 이원적 접근법에 해당하며, 데이터 시퀀스를 디스조인트 윈도우로 나누고, 질의 시퀀스를 슬라이딩 윈도우로 나눈다.

인덱스 구성단계에서 FRM 에서 발생했던 MBR 저장으로 인한 착오해답을 줄이기 위하여, 디스조인트 윈도우로 나누어진 데이터 시퀀스를 개별 점으로 직접 저장한다.

서브시퀀스 매칭단계에서 쿼리 시퀀스를 슬라이딩 윈도우로 나누고, 개별 점에 대해 범위 질의를 구성했던 FRM 과 달리 여러 점을 포함하는 1 개의 MBR 을 구성하여 범위 질의를 구성한다. MBR 에 포함된 각 점과 MBR 로 검색한 결과로 얻은 각 점 간의 f 차원 거리를 계산하여 $\varepsilon/\sqrt{p}$-매치하는 점들만 포함시키는 색인 수준 여과를 통해 최종 후보집합을 구한다.

위와 같은 과정을 통하여 이전의 연구보다 착오해답을 획기적으로 줄일 수 있었지만, 다양한 응용에서 사용하기에는 방법이 제한적이라는 문제점이 있다.

3.2.3 GeneralMatch

GeneralMatch 는 이전에 소개된 FRM 과 DualMatch 의 윈도우 구성 방법을 일반화한 것으로 FRM 의 알고리즘과 유사하다. 슬라이딩 윈도우와 디스조인트 윈도우를 일반화한 J-슬라이딩 윈도우와 J-디스조인트 윈도우를 이용하여, 기존의 인덱스 구성 및 서브시퀀스 매칭 단계를 수행한다.

FRM 은 J 값이 1 에 해당하는 경우이고 DualMatch 는 J 값이 ω에 해당하는 경우이다. 그러나 FRM 처럼 J 값이 작은 경우에는 색인에 저장하는 점의 수가 많아져 페이지 액세스 횟수 상에 문제가 생기기 때문에, 데이터의 종류와 크기, 사용되는 질의의 종류, 선택률의 범위 등에 따라 적합한 J 값을 찾아내는 방법을 제시하였다. 이로 인해 FRM 및 DualMatch 에 비해 보다 다양한 응용에서 적합하게 사용할 수 있는 방법이다.

4. 비교 분석

유사 시퀀스 매칭 방법은 기존의 매칭 방법을 일반화하는 형태로 발전되어왔다. 가장 처음 제안된 전체 매칭 방법은 데이터 시퀀스와 질의 시퀀스를 저차원 변환하여 각 점을 색인에 입력하기 때문에 저장 공간이 많이 들고 후처리 과정에 많은 시간이 소요된다. 전체 매칭 방법은 데이터 시퀀스와 질의 시퀀스가 동일한 길이를 가져야 하기 때문에, 동일하지 않더라도 매칭할 수 있는 방법인 서브시퀀스 매칭 방법이 연구되었다.

서브시퀀스 매칭 방법의 대표적인 연구인 FRM 은 데이터 시퀀스를 MBR 로 변환하여 색인화시켜 저장 공간을 줄일 수 있었다. 그러나 MBR 로 색인화시킴으로 인해 착오해답이 많이 발생하게 되었다. 이러한 문제를 해결해 후보집합을 줄임으로써 후처리 시간을 줄이는 DualMatch 가 연구되었다. 기존에 연구된 서브시퀀스 매칭 방법이 윈도우 구성에 있어 특정한 상황만 반영한 것이기 때문에, 이를 일반화하기 위한 방법이 GeneralMatch 이다. GeneralMatch 는 윈도우 구성을 자유롭게 바꿀 수 있어 다양한 응용에 따라 적합하게 사용할 수 있는 방법이다.

5. 결론

시계열 데이터는 다양한 분야에서 응용되고 있으며 그 필요성이 더욱 높아지고 있다. 본 논문에서는 시계열 데이터베이스 어플리케이션의 중심인 유사 시퀀스 매칭 방법에 대하여 조사하였다. 유사 시퀀스 매칭 방법을 전체 매칭과 서브시퀀스 매칭으로 나누어 살펴보고 대표적인 방법인 Agrawal 등의 전체 매칭 방법, FRM, DualMatch, GeneralMatch 에 대해 조사하고 분석하였다. 각 방법은 주로 윈도우 구성법, 크기 및 개수, 질의 시퀀스에 초점을 맞추어 발전되어 온 것을 알 수 있다. 이러한 문제점 외에 MBR 구성법, 전처리 변환 등 다양한 방법을 이용하여 효율적인 유사 시퀀스 매칭법에 관한 연구를 할 수 있을 것이다.

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] R. Agrawal, C. Faloutsos and A. Swami, Efficient similarity search in sequence databases. Proceedings of the International Conference on Foundations and Data Organization and Algorithm, 1993.

[3] 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.

[4] 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.

[5]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.