시계열 데이타의 유사성 검색을 위한 히스토그램 비교법

임동혁°, 김창룡, 정진완
한국과학기술원 전산학과

Histogram Comparing Technique for Similarity Search in Time-Series Data

Tong h. Lim°, Chang-Ryong Kim, Chin-Wan Chung
Department of Computer Science, KAIST

요 약

데이타웨어하우스의 주된 용도는 비지니스 의사결정이며, 이를 위한 경향 및 패턴을 찾는 문제는 매우 중요한 연구분야이다. 경향 및 패턴은 데이타웨어하우스 내의 데이타간의 상호관계를 분석함으로써 찾을 수 있는데, 이를 위한 유사성 검색기법 중 특히 뛰어난 3가지 기법들을 자세히 알아보고, 이들에 모두 적용 가능한 히스토그램 비교법을 제안하였다. 제안된 히스토그램 비교법을 이용하면 유클리디안 거리측정의 부담을 대폭 줄여, 전체 처리시간을 비약적으로 감소시킬 수 있다

1. 서론

데이타웨어하우스에 저장되는 데이타 중 상당수는 시간적 관계를 갖는 데이타의 집합이고, 이를 시계열 데이타(Time-Series Data)라 하는데, 경향이나 패턴은 이러한 시계열 데이타 간의 상호관계를 분석함으로써 찾을 수 있다. 최근 몇 년간 시계열 데이타의 저장 및 분석에 대한 작업 및 연구가 활발히 진행되고 있는데, 예를 들어, 금융회사들은 특정 기간이나 날짜의 주식 최고가, 최저가, 마감가, 거래량 등을 저장하고 있으며, 정보기관들은 매시간마다의 위성관련 데이타를 처리하고 있다. 이 외에도 제조업분야에서는 조립라인에서 매시간 발생하는 사건을 기록하고 있으며, 언론사들은 매일매일의 뉴스를 저장하고 있다. 또한 지구과학이나 공학분야에서는 지진파를 분석하거나 코드변환 등을 추적하는데 시계열 데이타를 사용하고 있다. 따라서 데이타웨어하우스에 저장된 시계열 데이타를 정확하고, 효율적으로 융통성 있게 다루는 것은 비지니스 성공에 극히 중요한 요소이다. 다음과 같은 예에서 우리는 시계열 데이타의 유사성 검색(Similarity Search)의 중요성과 높은 응용력을 알 수 있었고, 이를 향상시키는 기법을 연구하게 되었다.

본 논문에서는 시계열 데이타의 유사성 검색에 대한 기존의 연구를 자세히 알아보고, 기존 기법과 함께 사용되어 검색 시간을 비약적으로 향상시킬 히스토그램 비교법을 제안하고자 한다. 히스토그램 비교법은 사용자가 입력한 질의에서 얻은 히스토그램과 선행처리(Preprocessing)된 데이타웨어하우스 내의 시계열 데이타의 히스토그램을 직접 비교함으로써, 비교 대상의 수를 대폭 줄일 수 있으며, 이에 따라 전체 검색시간을 크게 줄일 수 있다.

2. 관련연구

[1]에서는 데이타웨어하우스 내의 모든 시계열 데이타 시퀀스들과 질의 시퀀스가 같은 길이를 가지고 있고, 각 시퀀스들은 N 차원 공간의 포인트로 간주되었다. 만일 두 포인트의 유클리디안 거리(Euclidean distance)가 임계치 ε 보다 작다면, 두 포인트에 해당하는 두 시퀀스는 유사하다고 간주하였다. 각 스퀀스가 하나의 포인트로 매핑되므로, R* 트리를 색인구조로 사용하여 포인트들을 저장하였다. 시퀀스들은 각각 f개의 특성을 사용하는 f 차원의 포인트들로 표현된다. 특성추출에는 이산 퓨리에 변환(Discrete Fourier Transform, DFT)을 사용하였는데, DFT를 사용한 이유는 DFT가 유클리디안 거리를 보존하기 때문이다. 변환 후 f개의 데이타가 시퀀스를 나타내는데 사용된다. 그림1은 [1]의 내용을 그림으로 나타낸 것이다. 사용자가 입력한 시계열 데이타 시퀀스 Q도 앞에서 설명한 과정을 거쳐 f 차원의 포인트로 변환되고, f 차원의 공간에서 Q와 유클리디안 거리가 임계치 ε 보다 작은 모든 시퀀스를 찾을 수 있다.

[3]에서는 [1]의 연구를 확장하여 사용자가 입력하는 질의 Q를 포함하는 시퀀스를 찾는 문제를 해결하였다. [1]은 데이타웨어하우스 내의 모든 시퀀스와 질의 시퀀스가 같은 길이를 가진다고 가정하므로, 길이가 다른 시계열 데이타 시퀀스를 비교하는데는 부적절하였다. [3]에서도 [1]에서와 마찬가지로 시퀀스의 특성을 추출하기 위해 DFT를 사용하였으니, 길이가 다른 시퀀스를 비교하기 위해 각 시퀀스를 가능한 모든 위치로부터 시작하는 ω 크기의 슬라이딩 윈도우(sliding window)들로 나눈 후, 슬라이딩 윈도우들을 서로 비교하였다. 따라서 시퀀스의 특성을 하나의 포인트로 나타내는 대신, 특성공간의 포인트들의 트레일(trail)로 나타내게 된다. 이러한 포인트들을 색인에 올리기 위해, 트레일들은 서브트레일들로 나누고, 각 서브트레일들은 최소 경계 사각형(Minimal Bounding Rectangle, MBR)으로 표현한다. 사용자 질의 Q도 역시 같은 과정을 거쳐 각 트레일을 포함하는 MBR이 구해진다. 이 질의 Q의 MBR과 교차하는 MBR을 갖는 시퀀스가 질의와 유사하다고 판단된다. 그림2는 임의의 시퀀스

S와 질의 Q를 ω 크기의 슬라이딩 윈도우로 나누는 것을 보인 것이다

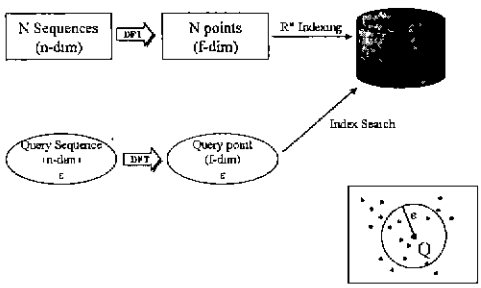
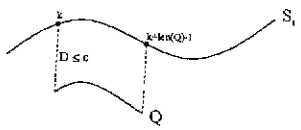


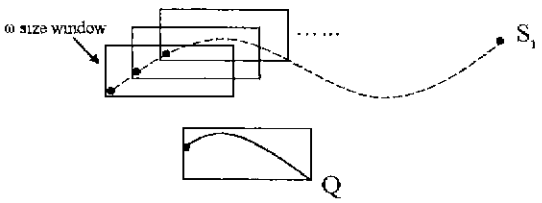
그림1 유클리디안 거리측정을 통한 유사성 검색

[2]는 잡음이나 스케일링(scaling) 또는 이행이 있는 시계열 데이타들 간의 유사성 검색을 가능하게 해준다. [2]에서는 두 시퀀스가 유사한 서브시퀀스의 쌍을 충분히 가지고 있을 때, 두 시퀀스는 유사하다고 본다. 이 때, 서브시퀀스의 쌍들은 겹치지 않는 시간 순차적 쌍들(non-overlapping time-ordered pairs)이어야만 한다. 두 시퀀스를 비교하는 과정에서 시퀀스의 일부분은 아웃라이어(outlier)로 간주되어 제거될 수 있으며, 두 시퀀스의 시간축에 반드시 정렬되어 있을 필요는 없다. 유사성 비교는 한 시퀀스로부터 임계치 ε 이내에 다른 시퀀스가 있는지 검사함으로서 행해진다. 물론 아웃라이어는 무시된다. 그림3은 [2]의 유사모델을 보인 것이다 그림3의 (1)의 시퀀스 S와 시퀀스 T가 유사한 지를 검사하기 위해 그림3의 (2)와 같이 S의 아웃라이어는 제거될 수 있다. 또한, 그림3의 (3)과 같이 T의 오프셋을 조정하여 두 시퀀스의 높이를 맞출 수 있으며, 그림3의 (4)에서처럼 진폭변환을 통해 스케일링을 할 수도 있다 이러한 변환을 거친 S와 T가 임계치 ε 이내에서 겹쳐지면 두 시퀀스는 유사하다고 할 수 있다.

Find S_i such that $D(Q, S[k:k+len(Q)-1]) \le \varepsilon$



(a) 서브시퀀스 Q가 시퀀스 S가 유사한 조건



(b) ω 크기의 슬라이딩 윈도우로 나뉜 시퀀스

그림2 슬라이딩 윈도우 기법

3. 히스토그램 비교법

기존의 유사성 검색기법[1][2][3]은 모두 시계열 데이타 시퀀스를 DFT 등의 신호처리기법을 이용하여 다차원공간 내의 하나의 포인트로 변환한 후, 포인트간의 유클리디안 거리를 비교하는 방법을 사용하였다. 따라서 질의 시퀀스에서 얻은 포인트와 데이타웨어하우스 내의 모든 시퀀스에서 얻은 포인터들 간의 거리를 비교하는데 상당한 부담이 있는 것이 사실이다 유클리디안 거리측정은 모두 다차원공간에 행해지며, 또한 모두 질의시간에 실행하여야 하므로 그 부담은 상당하다고 하겠다 이 때문에 다차원 유클리디안 거리측정보다 처리속도가 빠르면서, 될 수 있으면 신행처리(preprocessing) 시간에 대부분의 작업을 행하여 질의시간의 부담을 감소시키는 기법에 대한 연구가 필요하였다.

히스토그램 비교법은 신행처리시간에 데이타웨어하우스의 각 시퀀스의 히스토그램을 미리 구해 놓은 후, 질의 시퀀스의 히스토그램과 이들을 비교하는 기법이다 두 개의 히스토그램을 비교하는 것은 단순비교 연산으로 시간 복잡도가 O(1)이며 데이타웨어하우스 내의 모든 시퀀스와 비교하는데도 O(n) 밖에 걸리지 않는다. 실제 이 기법을 사용하면, 대부분의 처리시간을 히스토그램 생성에 쓰이게 되는데, 이는 선행처리시간에 해당되며, 질의시간에는 질의 시퀀스 하나에 대해서만 히스토그램을 생성하므로, 그 부담이 미미하다고 할 수 있다.

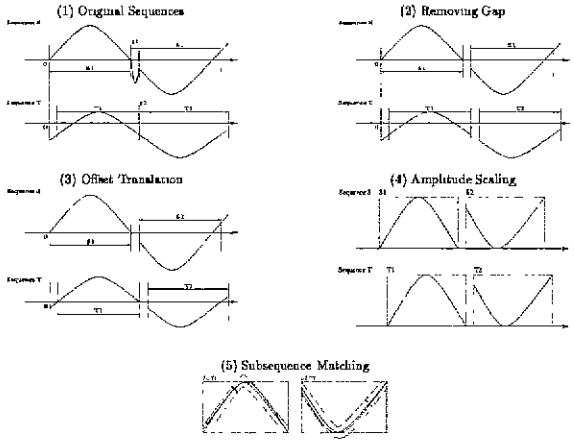


그림3 잡음, 스케일링, 이행이 있는 환경에서의 시퀀스 비교

정의1: n개의 데이타로 이루어진 시계열 데이타 시퀀스를 S라고 하면, S는 $(S_1, S_2, S_3, .., S_n)$이다.

정의2: 시계열 데이타 시퀀스 S 내의 i번째 위치에 있는 원소를 S[i]로 표기한다.

정의3: 시계열 데이타 시퀀스 S내의 i번째 위치로부터 j번째 위치까지의 원소로 구성된 서브시퀀스를 S[i, j]로 표기한다.

정의4: 서브시퀀스 S[i, j]의 길이는 j - i + 1이 된다.

정의5: 시계열 데이타 시퀀스 Q가 시계열 데이타 시퀀스 S의 서브시퀀스 S[i, j]와 같거나 임계치 ε 내에서 비슷하면, Q와 S는 유사하다고 한다.

정의6: 시계열 데이타 시퀀스 S의 각 원소의 빈도수를 구하여 생성된 히스토그램을 H(S)로 표기한다.

히스토그램 비교법이 가능한 이유는 시계열 데이타 시퀀스 S의 히스토그램 H(S)와 시계열 데이타 시퀀스 Q의 히스토그램 H(Q)가 있을 때, H(S)의 모든 원소 S_i에 대해 H(Q)의 모든 원소 Q_i가 작거나

같을 때, H(S)에 H(Q)가 포함되기 때문이다

정리1(히스토그램의 포함관계): 시계열 데이타 시퀀스 Q가 시계열 데이타 시퀀스 S에 포함되어 있을 때, Q의 히스토그램 H(Q)는 S의 히스토그램 H(S)에 포함된다 즉 $H(Q) \subseteq H(S)$이다.

증명: 시계열 데이타 시퀀스 S에서 시계열 데이타 시퀀스 Q와 일치하는 서브시퀀스를 제거하여 만들어지는 시퀀스를 T라하고, $H(T)=<T_1, T_2, T_3, ., T_n>$, $T_i \geq 0$ 이라 하면, $H(S)=<S_1, S_2, S_3, .., S_n>$, $H(Q)=<Q_1, Q_2, Q_3, .., Q_n>$ 일 때, $S_i = T_i + Q_i$이므로 모든 i에 대하여 $S_i \geq Q_i$이다. 따라서 $H(Q) \subseteq H(S)$이다

끝(정리1)

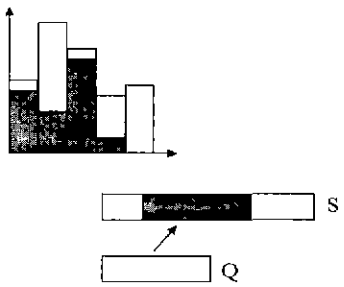


그림4 히스토그램의 포함관계

실제로 두 히스토그램 H(S)와 H(Q)를 비교하기 위해선, H(S) 내의 H(Q)에 해당하는 부분과 비교해야 하므로, H(Q)와 $H(S \circ Q)$가 같은 지를 비교한다 $H(S \circ Q)$는 두 히스토그램 H(S)와 H(Q) 중 작은 값으로만 이루어진 히스토그램을 의미한다. H(Q)와 $H(S \circ Q)$가 같거나 일정한 범위에서 흡사하면, S와 Q는 유사할 가능성이 높은 것으로 보고, [1][3][2]에서 사용한 방법을 적용하여 유클리디안 거리를 측정한다. 만일 H(Q)와 $H(S \circ Q)$가 전혀 다르거나 허용하는 한계를 넘는 차이를 보이면 S와 Q는 유사할 가능성이 없는 것으로 보고 유클리디안 거리측정의 대상에서 제외시킬 수 있다.

정의7: 시계열 데이타 시퀀스 S와 시계열 데이타 시퀀스 Q의 히스토그램이 각각 $H(S) = <S_1, S_2, S_3, ..., S_n>$, $H(Q) = <Q_1, Q_2, Q_3, , Q_n>$ 일 때, 히스토그램 $H(S \circ Q)$는 다음과 같이 정의된다.

$H(S \circ Q) = <\min(S_1, Q_1), \min(S_2, Q_2), \min(S_3, Q_3), ..., \min(S_n, Q_n)>$

예1과 예2는 히스토그램 비교법을 통해 두 시퀀스가 유사할 가능성이 있는지 여부를 결정하는 과정을 보인 것이다

예1(유사 가능한 시퀀스): 데이타웨어하우스에 저장되어 있는 시계열 데이타 시퀀스 S에 대해 사용자가 입력한 질의 시퀀스 Q가 유사한지 알고자 한다. S와 Q는 각각 다음과 같다

S = (1, 2, 3, 5, 2, 3, 4, 5, 1, 3, 2, 4)

Q = (2, 3, 4, 5, 2, 5, 2, 4)

선행처리로 구해져있는 S의 히스토그램 H(S)와 질의 Q로부터 얻은 히스토그램 H(Q)는 다음과 같다.

$H(S) = <2, 3, 3, 2, 2>$

$H(Q) = <0, 3, 1, 2, 2>$

H(S)와 H(Q)를 비교하여 구한 $H(S \circ Q)$는 다음과 같다.

$H(S \circ Q) = <0, 3, 1, 2, 2>$

이 $H(S \circ Q)$와 질의 Q로부터 얻은 히스토그램 H(Q)를 비교하면, 두 시계열 데이타 시퀀스 S와 Q가 유사할 수 있음을 알 수 있다.

끝(예1)

예2(유사 불가능한 시퀀스): 데이타웨어하우스에 저장되어 있는 시계열 데이타 시퀀스 S'에 대해 사용자가 입력한 질의 시퀀스 Q가 유사한지 알고자 한다. S'과 Q는 각각 다음과 같다

S' = (2, 3, 3, 3, 3, 4, 5, 3, 3, 3, 1, 1)

Q = (2, 3, 4, 5, 2, 5, 2, 4)

선행처리로 구해져있는 S'의 히스토그램 H(S')와 질의 Q로부터 얻은 히스토그램 H(Q)는 다음과 같다.

$H(S') = <2, 1, 7, 1, 1>$

$H(Q) = <0, 3, 1, 2, 2>$

H(S')와 H(Q)를 비교하여 구한 $H(S' \circ Q)$는 다음과 같다.

$H(S' \circ Q) = <0, 1, 1, 1, 1>$

이 $H(S' \circ Q)$와 질의 Q로부터 얻은 히스토그램 H(Q)를 비교하면, 두 시계열 데이타 시퀀스 S'과 Q가 유사할 가능성이 없음을 알 수 있다.

끝(예2)

4. 결론

데이타웨어하우스에 저장된 시계열 데이타간의 유사성을 검색하는 기법들 중 특히 뛰어난 3가지 기법에 대해 알아보았고, 이들 기법에 적용 가능한 히스토그램 비교법을 제안하였다. 히스토그램 비교법은 사용자가 입력한 질의와 데이타웨어하우스 내에 저장된 시퀀스가 유사가능한가를 빠른 속도로 검사하여 돌려줌으로써, 다치원 유클리디안 거리측정의 횟수를 대폭 줄일 수 있으며, 히스토그램 생성 등의 대부분의 처리를 선행처리시간에 행할 수 있으므로 질의시간에 부담을 주지 않는다. 또한 히스토그램 비교법은 기존 연구들이 처리하지 못했던, 다차원 유사성 검색에도 응용될 수 있다. 다차원 시계열 데이타에 대한 유사성 검색은 비디오 데이타를 포함하여 밀티미디어, 지리정보시스템 등의 대용량, 비정형 데이타 검색 분야에서 응용될 수 있으므로 그 파급효과가 매우 크다.

참고문헌

[1] R. Agrawal, C. Faloutsos,and A. Swami. Efficient similarity search in sequence databases In Proc. 4th Intl Conf on Foundations of Data Organization and Algorithms, pages 69-84, October 1993.

[2] R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In Proc 21st Int. Conf Very Large Data Bases, pages 490-501, Zurich, Switzerland, Sept. 1995.

[3] C Faloutsos, M Ranganathan, and Y Manolopoulos Fast subsequence matching in time-series databases. In Proc. ACM SIGMOD Int. Conf. Management of Data, pages 419-429, Minneapolis, May 1994