시계열 데이타베이스에서의 트렌드 유사도 탐색

이지은 °, 윤종필
숙명여자대학교 전산학과 데이타베이스 연구실

Trend Similarity Search In Time-Series Databases

Jieun Lee °, Jongpil Yoon
Database Lab, Dept of Computer Science, Sookmyung Women's University

요약

최근 시계열 데이타에서 유사한 패턴을 탐색하는 기법이 다양한 응용분야에서 중요한 연구 주제로 자리잡고 있다. 본 논문에서는 시계열의 트렌드를 정의하고 유사한 트렌드를 가진 시계열을 찾음으로써 유사성의 개념을 좀 더 확장, 발전 시켰다. 즉, 시계열에서의 트렌드를 두 개의 이동 평균 선의 관계를 통해 정의함으로써 두 시계열 간의 거리만으로 유사도를 측정했던 기존 연구와는 달리 좀 더 패턴 변화에 의미있게 접근하여 유사도를 탐색할 수 있다. 두 개의 이동 평균 선의 관계 속에서 트렌드 기호를 정의하고 이 기호가 해당되는 범위를 같이 정의해주는 것을 통해서 시계열의 트렌드를 대표하게 되며, 이것을 데이타베이스에 저장, 검색하여 트렌드 탐색을 한다.

이러한 개념으로 시계열간에 유사 패턴을 가진 수열들을 찾고 이것을 기존의 DFT 방법을 이용하여 대용량의 시계열 데이타베이스에서 사용자가 정의한 임계치 이하로 차이가 나는 시계열에 대해 유사 시계열로서 최종적으로 검색하게 된다.

1. 서론

시계열 데이타는 주식, 기상, 지리, 천체 물리학, 의학 등의 여러 분야에서 자연스럽게 발생되며, 최근 시계열 데이타베이스에서 유사 패턴(similarity pattern)을 탐색하는 기법이 광범위한 자연과학 및 비즈니스 응용분야에서 중요한 연구분야로 여겨지고 있다 [1,2,3,4,5,6,7,8,9]. 기존에는 시계열 데이타베이스 상에 정확히 같은 결과를 요구하는 질의를 주로 사용한 반면, 최근에는 어느 정도 서로 유사한 결과를 요구하는 질의문이 많이 사용되고 있다. 예를 들면[1], 유사한 패턴으로 성장하는 기업체를 검색한다거나, 판매 패턴이 서로 유사한 상품을 검색, 또는 주식 시세의 이동이 유사한 분포를 보이는 것을 찾으라는 등의 질의와 같다.

일반적으로 시계열 Q 는 등간격으로 발생되어진 실수들의 연속된 집합 $(q_1, \ldots, q_n)$ 을 뜻하며, q_i 는 어느 한 시점에서의 값을 나타낸다. 현재 시계열 탐색에 있어서 주로 초점을 두는 문제는 다음과 같다 [4].

질의 수열 (query sequence) 을 $Q = (q_1, \ldots, q_n)$, 데이타 수열들의 집합을 $S = (S_1, \ldots, S_t)$ 라고 할 때, 길의 수열 Q 와 유사한 S_i 의 부분 수열 S' 을 찾아내는 것이다 $(1 \le i \le l)$. 유사도를 측정하는 기준은 보통 L_p 미터법의 거리 함수를 사용한다. 즉, 두 개의 수열 $X = (x_1, \ldots, x_n)$ 와 $Y = (y_1, \ldots, y_n)$ 가 있을 때, 이들 간의 거리 $D_p(X,Y)$ 는 다음과 같다.

$$D_p(X, Y) = \sqrt[p]{\sum_{i=1}^{n} | x_i - y_i |^p} \qquad \text{<식 1>}$$

그리고 시계열 유사도 측정에 있어서는 대부분의 경우 $p = 2$, 즉 유클리드 거리를 이용하여 측정한다.

본 논문에서는 시계열간의 유사성을 탐색할 때, 단순히 두 시계열간의 거리상의 근접도만을 측정하여 유사도를 결정하는 것이 아니라 트렌드라는 개념을 새롭게 적용하여 더욱 의미 있는 결과 집합을 도출하는데 그 초점을 둔다.

본 논문의 구성은 다음과 같다. 먼저 제 2절에서는 관련 연구를 요약하고, 제 3절에서는 이산 푸리에 변환의 정의 및 특성에 대하여 간단히 설명한다. 제 4절에서는 트렌드의 정의와 이를 위한 트렌드 기호를 설명하고 사용 예제를 제시한다. 제 5절에서는 본 논문에서 사용한 탐색 시스템의 알고리즘을 요약하였고 제 6절에서는 결론을 맺는다.

2. 관련 연구

시계열 데이타에서 유사성에 기반을 둔 매칭은 이미 신호 처리 분야, 특히 음성 처리 영역에서 많이 연구해오고 있었다. 그러나 대량의 데이타 집합에서의 효율적 탐색보다는 작은 양의 데이타 집합에서의 정확성을 더 중시했다.

대량의 데이타베이스에서 무엇보다 중요한 요소는 바로 속도이고, [1]에서는 동일 길이의 시계열 데이터에서의 유사성 탐색 기법으로 이산 푸리에 변환 (Discrete Fourier Transform (DFT)) 을 이용했다. 이것은 각 데이타 수열을 첫 번째의 DFT 계수 즉 하나의 점으로 대체하여 R^*-tree 를 이용한 F-$Index$ 를 만드는 것을 제안한다. 그리고 질의 수열과의 거리 차가 사용자가 결정한 오류 거리 ε 보다 작은 수열을 R^*-tree 에서 탐색한다. 이 방법은 거짓 삭제 (false dismissal) 가 없다는 것을 보장하지만, 거짓 추출 (false alarm) 은 발생할 수 있다.

[1]을 좀더 일반화시킨 것으로서, 서로 다른 길이의 데이타 수열들간의 유사성 탐색을 위해 ST-$index$ 를 제안한 방법이 있다 [5]. n 크기의 최전 윈도우 (sliding window) 가 데이타 수열들 상의 모든 가능한 위치를 이동하면서 각 윈도우 안의 부분 수열들을 n깊의 DFT 를 이

용하여 변환시키는 것이다 이러한 변환을 통해서 각 개수들의 흔적(trail) 이 험성되고, 이 흔적들은 최소 바운드 사각형(minimum bounding rectangle (MBR)) 을 이용하여 부분 흔적들(subtrails) 로 나뉘어 R^*-tree 에 저장된다. 범위성 질의(range query) 의 경우 질의 지역과 교차되는 모든 MBR 을 검색하는 것이다. 이 방법 역시 거짓 삭제가 없음을 보장하지만, 거짓 추출의 가능성을 가지고 있으며 이를 위해 후처리 (post-processing)를 요한다.

위의 두 방법들은 유사성을 탐색할 때 대상 수열들의 어떤 변환도 고려하지 않는다 그래서 좀 더 유사 개념을 확장한 것이 나오는데, 각 데이타 수열들을 적당량의 비율로 확대 또는 축소하거나 적당량의 오프셋만큼 변형하여 유사하게 변환된다면 변환된 수열들이 서로 유사하다고 생각하는 것이다 [2]. 여기서는 서로 매칭 되지 않는 갭 또한 허용한다

유사 개념의 확장이라는 점에서 같은 맥락에 있다고 볼 수 있는 것으로서, 이동 평균(moving average)의 방법을 제안한 경우도 있다 [8]. 기존 주식 분석에서 많이 사용되는 방법인데, 5일 또는 20일 등의 이동 평균을 이용하여 원래의 수열을 좀 더 완만하게 평활화(smoothing) 하는 것이다 이렇게 변환된 수열들을 대상으로 유사도를 탐색하므로 좀 더 많은 결과 집합을 얻을 수 있게 된다.

또한 대량의 데이타베이스에서 원하는 수열을 찾는 것이 시간을 요하므로 유사하지 않은 수열을 좀 더 빨리 인식하므로 시간을 단축시키는 방법, 즉 lower bounding 기법을 제안한 경우도 있다 [9]. 여기에서는 기존의 DFT 를 이용하기 보다 FastMap을 가지고 수열들을 색인하며, 시간축의 데이타를 늘이거나 줄임(time warping)으로 유사도의 범위를 확장시켜 적용하고 있다.

DFT 변환 외에 웨이블릿 변환(wavelet transform)을 이용하여 효과적인 탐색 기법과 시계열 데이타의 차원을 감소시키는 방법을 제안하기도 하였다 [3].

3. 이산 푸리에 변환 (The Discrete Fourier Transform)

시계열 데이타는 대부분 길고, 이들간의 거리 계산에는 많은 시간이 소요된다 이를 위한 해결책으로 시간 영역의 수열을 주파수 영역으로 변환시켜주는 푸리에 변환법을 사용할 수 있다 [1] 본 절에서는 이산 푸리에 변환의 특징에 대해 간단히 살펴보고자 한다.

먼저 $i = 0, 1, .., n-1$ 에 대한 유한 길이의 신호 $\vec{x} = [x_i]$ 에 대해서 n 개의 점의 이산 푸리에 변환은 $F = 0, 1, .., n-1$ 에 대한 n 개의 복소수 X_F 의 수열이라고 한다면 X_F 는 다음과 같다

$$X_F = 1/\sqrt{n} \sum_{t=0}^{n-1} x_t \exp(-j\,2\pi Fi/n) \quad F = 0, 1, \ldots, n-1$$

<식 2>

여기서 j 는 허수 부분으로 $j = \sqrt{-1}$ 이다. 신호 $\vec{x}$ 는 역변환으로 다음과 같이 구해질 수 있다.

$$x_t = 1/\sqrt{n} \sum_{F=0}^{n-1} X_F \exp(j\,2\pi Fi/n) \quad t = 0, 1, \ldots, n-1$$

<식 3>

여기서 X_0를 제외한 X_F 는 복소수이다 단 신호 $\vec{x}$ 는 실수이다

그리고, 수열 $\vec{x}$ 의 에너지는 수열의 각 시점의 에너지들의 합으로 나타난다.

$$E(\vec{v}) \equiv \|\vec{x}\|^2 \equiv \sum_{i=0}^{n-1} |x_i|^2$$

<식 4>

또한 Parseval의 정리로 인해서 DFT는 신호의 에너지를 보존한다라는 것을 보일 수 있다.

[정리(Parseval)] 수열 $\vec{x}$ 의 DFT를 $\vec{X}$ 이라 두면 다음과 같은 식을 얻을 수 있다

$$\sum_{t=0}^{n-1} |x_t|^2 = \sum_{F=0}^{n-1} |X_F|^2$$

<식 5>

그리고, DFT는 선형 변환이기 때문에 DFT는 두 개의 신호 $\vec{x}$, $\vec{y}$ 간의 유클리드 거리를 보존한다는 것을 보일 수 있게 된다

$$D(\vec{x}, \vec{y}) = D(\vec{X}, \vec{Y})$$

<식 6>

여기서 $\vec{X}$, $\vec{Y}$ 는 $\vec{x}$, $\vec{y}$ 의 푸리에 변환이다.

한편 DFT의 계수 중 앞의 몇 개(2-3)만을 취하고, 나머지는 버리기 때문에 실제 두 수열간의 거리보다 더 작은 값으로 나오게 되며, 이것은 거짓 삭제가 없음을 보장하게 된다

또한 이 방법은 수열을 탐색할 때 보다 적은 데이타, 즉 몇 개의 푸리에 계수 값으로 색인을 하여 수행하므로 좀 더 작은 공간에서 빠르게 탐색할 수 있게 해준다.

그러나 이러한 접근은 사용자가 오류 거리를 지정하는 것 외에 유사성의 의미를 제어할 수 없으며, 좀 더 의미 있는 시계열 분석을 하기에 적합하지 않다 그러므로 본 논문에서는 시계열의 트렌드라는 개념을 새롭게 도입하여 먼저 트렌드의 유사성을 검색하는 것을 제안한다

4. 트렌드의 정의

보통 시계열의 트렌드를 분석하는 분야는 주식 데이타의 경우이며 이에 대한 연구는 활발히 일어나고 있다 [7] 기존의 트렌드를 위해 사용한 방법은 이동 평균법이지만, 이것은 단순히 대략적 경향을 손쉽게 나타내주는데 그치고 있다. 트렌드는 두 개의 이동 평균이나 트렌드 선을 이용하여 정의할 수 있으며, 이것은 시계열 패턴의 변화를 보여주는데 이용될 수 있다 이를 위해 여섯 가지 기호가 사용되며 이들의 상호작용을 통해서 트렌드를 분석 및 탐색할 수 있다

4.1 트렌드 기호의 정의

먼저 시점 $t = 0, 1, .., n-1$에 대하여 m주 이동 평균 선을 $MA_m(t)$, k주 이동 평균 선을 $MA_k(t)$ 그리고 실제 시계열 데이타를 $X(t)$ 라고 할 때 트렌드 기호의 각 정의는 다음과 같다 (단, m < k)

▲ , △ (상승세)

$MA_m(t)$ 이 $MA_k(t)$ 보다 클 경우

▼ , ▽ (하락세)

$MA_m(t)$ 이 $MA_k(t)$ 보다 작을 경우

■ (교차 지점)

$MA_m(t)$ 이 $MA_k(t)$ 위로 교차되는 경우

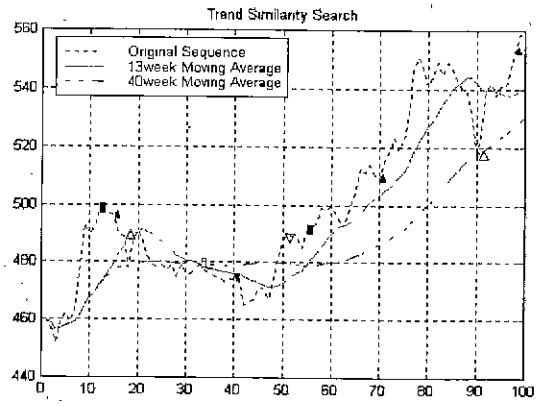
□ (교차 지점)

$MA_m(t)$ 이 $MA_k(t)$ 아래로 교차되는 경우

단기간의 트렌드는 주로 m주 이동평균이 나타내며, 트렌드는 m주 이동평균의 3% 범위를 계산하여 구하게 된다. 즉, 상승세와 하락세의 기호 중 △ , ▽ 는 실제 시계열 데이타가 $MA_m(t)$ 의 3%범위를 넘었음을 나타낸다 이것은 기존 트렌드가 점차 약화됨을 반영한다 반면, ▲ , ▼ 는 실제 시계열 데이타가 여전히 장기간의 트렌드와 동일하며 $MA_m(t)$ 의 3%범위를 넘지 않는다는 것을 뜻한다 <표 1>은 트렌드 기호를 정의한 것이다.

4.2 예제

IBM사의 주식 시세 데이타를 가지고 13주 이동 평균선과 40주 이동 평균선간의 관계를 통하여 4.1절의 정의에 따라 트렌드를 구한 결과는 다음과 같다 [10].



< 그림 1 > IBM사의 주식 시세 트렌드

두 개의 수열에서 위에서 정의한 트렌드가 같으면 수열간의 거리 차를 비교하기에 앞서 유사한 시계열로 간주한다. 즉, 트렌드의 기호와 범위가 사용자가 정의한 임계치만큼의 차이기 있다면 트렌드가 유사하다고 보고 유사개념을 보다 의미있게 해준다

5. 알고리즘

질의 수열을 가지고 먼저 트렌드를 찾아내고, 각 시계열 데이타들의 트렌드가 유사한 것들을 찾아낸다 이것을 통해 걸러진 각 시계열 데이타들에 대해서 DFT로 변환하고 이 DFT 계수를 가지고 R*-tree를 통해 인덱스를 만들고 이를 통해 사용자가 원하는 임계치 만큼의 차이를 두는 수열들을 찾아낸다

```
  O(t) ← 데이타 수열들 , q ← 질의 수열
( t = 0, 1, ,n-1 )
q_trend = TREND(q)
for i ← 0 to n-1
 do O_trend(i) = TREND( O(i) )
    if O_trend(i) - q_trend < ε
      then trend_result ← O_trend(i)
/* 아래 부분은 선택적으로 수행될 수 있다
 O_index = create_index(DFT(trend_result))
FIND( DFT(q), O_index)
```

< 유사 수열 탐색 알고리즘 >

6. 결론

기존의 연구에서는 수열간의 거리로써 유사도를 측정하였지만, 이것은 이미 DFT 변환 후의 적용이기 때문에 좀더 확장된 유사개념을 세우기가 쉽지 않았다. 물론 시계열을 확장 또는 축소, 역회전 그리고 이동평균에 의한 평활화등 적용시켰지만, 거리의 한계 안에서 유사개념을 결정 지을 수밖에 없었다 그러나 기존 주식 데이타의 트렌드 개념을 적용함으로써 시계열간의 트렌드를 발견할 수 있게 되었다

본 논문에서는 시계열 데이타베이스에서 질의 수열과 유사한 데이타 수열을 탐색하는 과정에서 트렌드라는 개념을 적용하여 유사도의 범위를 좀 더 확장 발전시킨 데 이 논문의 의의가 있다.

향후 연구 과제로는 트렌드라는 개념을 좀더 재계화 할 필요가 있으며, 이것을 검색하는데 좀더 빠르고 효과적인 방법 모색이 필요하다고 본다

참고문헌

[1] Rakesh Agrawal, Christos Faloutsos, and Arun Swami, Efficient similarity search in sequence databases, *the Fourth International Conference on Foundations of Data Organization and Algorithm, Chicago October* 1993

[2] Rakesh Agrawal, King-Ip Lin, Harpreet S Sawhney, and Kyuseok, Shim, Fast similarity search in the presence of noise, scaling, and translation in time series databases, *the VLDB Conference, Zurich, Switzerland, Sept* 1995.

[3] K. P. Chan and W. C Fu Efficient Time Series Matching by Wavelets *In International Conference on Data Engineering,* 1999

[4] Kelvin Kam Wing Chu, Man Hon Wong, Fast Time-Series Searching with Scaling and Shifting, *ACM PODS, Philadelphia PA, July* 1999

[5] C Faloutsos, M Ranganathan, and Y Manolopoulos, Fast subsequence matching in time-series databases, *the ACM SIGMOD Conference on Management of Data, May* 1994

[6] H. V Jagadish, A. O Mendelzon, and T Milo Similarity-Based Queries. *In Symposium on Principles of Database Systems, pages* 36-45, 1995.

[7] Skot Kortje, Stock Trends Analyst and Editor, Stock Trends -A Handbook for Investors, *http://www.stocktrends.ca/handbook_1.htm*

[8] Davood Rafiei and Alberto Mendelzon, Similarity-Based Queries for Time Series Data, *the ACM SGMOD Conference, Tucson, AZ, May* 1997.

[9] Byoung-Kee Yi, H. V Jagadish, and Christos Falutsos, Efficient Retrieval of Similar Time Sequences under Time Warping, *the 14th Int'l Conference on Data Engineering, Orlando, FL, Feburary* 1998.

[10] Makridakis, Forecasting- Methods and Applications (3rd edition) Data Sets, *http://www-personal.buseco.monash.edu.au/~hyndman/forecasting/ Wheelwright and Hyndman,* 1998

A. Appendix

기 호	정 의
▲	$MA_k(t) < MA_m(t)$, $MA_m(t) < X(t)$ or $MA_k(t) < MA_m(t)$, $MA_m(t) - MA_m(t) * 0.03 < X(t)$
△	$MA_k(t) < MA_m(t)$, $X(t) < MA_m(t) - MA_m(t) * 0.03$
▼	$MA_m(t) < MA_k(t)$, $X(t) < MA_m(t)$ or $MA_m(t) < MA_k(t)$, $X(t) < MA_m(t) + MA_m(t) * 0.03$
▽	$MA_m(t) < MA_k(t)$, $MA_m(t) + MA_m(t) * 0.03 < X(t)$
■	$MA_k(t) < MA_m(t)$
□	$MA_m(t) < MA_k(t)$

<표 1 > 트렌드 기호의 정의