정규 거리에 기반한 시계열 데이타베이스의 유사 검색 기법

(Similarity Search in Time Series Databases based on the Normalized Distance)

이 상 준 † 이 석 호 ††
(Sangjun Lee) (Sukho Lee)

요 약 본 논문에서는 정규 거리에 기반한 유사 시퀀스의 검색 기법을 제안한다. 시퀀스의 형태가 중요한 관심 사항인 응용에서 정규 거리는 단순한 Lp 거리에 비해 적합한 유사도라 할 수 있다. 이러한 정규 거리에 기반한 질의를 처리하기 위한 기존의 기법들은 시퀀스의 평균을 구한 후 이를 이용하여 시퀀스를 수직 이동하는 전처리 과정을 가지고 있다. 제안된 기법은 시퀀스의 인접한 두 요소들 간의 변이가 정규화 과정에 불변이라는 속성을 이용하여 수직 이동의 전처리 과정 없이 특징 벡터를 추출한 후 이를 R-tree와 같은 공간 접근 기법을 이용하여 인덱싱한다. 제안된 기법은 비슷한 형태의 시퀀스를 검색할 수 있으며 착오 누락이 없음을 보장한다. 실제 주식 데이타를 이용한 실험을 통해 제안된 기법의 성능을 확인하였다.

키워드 : 데이타베이스, 시계열, 유사 검색, 정규 거리

Abstract In this paper, we propose a search method for time sequences which supports the normalized distance as a similarity measure. In many applications where the shape of the time sequence is a major consideration, the normalized distance is a more suitable similarity measure than the simple Lp distance. To support normalized distance queries, most of the previous work has the preprocessing step for vertical shifting which normalizes each sequence by its mean. The proposed method is motivated by the property of sequence for feature extraction. That is, the variation between two adjacent elements of a time sequence is invariant under vertical shifting. The extracted feature is indexed by the spatial access method such as R-tree. The proposed method can match time series of similar shape without vertical shifting and guarantees no false dismissals. The experiments are performed on real data(stock price movement) to verify the performance of the proposed method.

Key words : database, time series, similarity search, normalized distance

1. 서 론

시간의 흐름에 따라 순차적으로 생성되는 데이타의 연속적인 모임인 시퀀스는 컴퓨터에 저장되는 데이타에 있어서 많은 부분을 차지하고 있다[1]. 이러한 시퀀스 중에서 숫자로 표현되는 시퀀스를 시계열이라 한다. 시계열 데이타베이스에서 유사한 시퀀스를 검색하는 것은 미래를 예측하고 가설을 검증하는 데이타마이닝[2,3]과 같은 분야에서 매우 중요한 역할을 하고 있다[4-7].

유사한 시퀀스를 효율적으로 검색하기 위한 많은 기법들이 단순한 Lp 거리에 기반하여 제안되어 왔다[1, 8-12]. 길이가 n인 두 시퀀스 $A=(a_1, a_2, \cdots, a_n)$와 $B=(b_1, b_2, \cdots, b_n)$ 사이의 단순한 Lp 거리는 다음과 같이 구할 수 있다.

$$L_p^{simple}(A,B) \equiv \left(\sum_{i=1}^{n} |a_i - b_i - m|^p \right)^{1/p}$$

여기서 $m = \sum_{i=1}^{n} (a_i - b_i)/n$

그러나 유사도로서 단순한 Lp 거리는 다음과 같은 문제점을 가지고 있다. 단순한 Lp 거리는 두 시퀀스의 절대적인 오프셋에 영향을 받는다. 따라서 비슷한 모양을 가진 시퀀스라 할지라도 수직적인 위치가 다르다면 유사하지 않다고 분류될 수 있다. 그림 1에 나타난 것같이

· 본 연구는 2003년도 두뇌한국 21 사업과 정보통신부의 대학 IT 연구센터(ITRC) 지원을 받아 수행되었습니다.
† 학생회원 : 서울대학교 전기컴퓨터공학부
freude@dbmain.snu.ac.kr
†† 종신회원 : 서울대학교 전기컴퓨터공학부 교수
shlee@cse.snu.ac.kr
논문접수 : 2003년 5월 24일
심사완료 : 2003년 10월 13일

질의 시퀀스 Q=(4 9 4 9 4)와 두 데이타 시퀀스 A=(7 5 6 7 6)와 B=(14 19 14 19 14)가 주어진 경우를 살펴보자. 단순한 Lp 거리를 유사도로 사용한다면 시퀀스 A가 시퀀스 B보다 질의 시퀀스 Q에 유사한 시퀀스로 판정된다. 그러나 형태를 살펴보면 시퀀스 B가 시퀀스 A보다 질의 시퀀스 Q에 유사하다고 볼 수 있다. 이러한 예에서 알 수 있듯이 시퀀스의 모양이 관심 사항인 응용에서는 단순한 Lp 거리가 적합한 유사도가 아님을 알 수 있다.

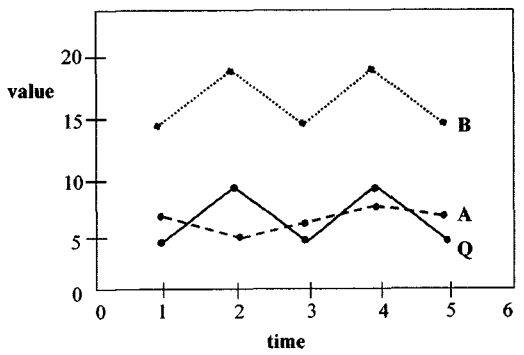


그림 1 단순 Lp 거리의 단점

이러한 단순한 Lp 거리의 단점을 해결하기 위해서 정규 거리가 두 시퀀스 간의 유사도로 종종 사용되어 왔다[5,13]. 정규 거리는 두 시퀀스의 오프셋을 무시한 거리로 시퀀스의 수직적 위치에 관계없이 모양이 유사한 시퀀스를 검색하는 경우에 적절한 유사도로 사용될 수 있다. 길이가 n인 두 시퀀스 A=(a_1, a_2, ···, a_n)와 B=(b_1, b_2, ···, b_n) 사이의 임의의 Lp에 대한 정규 거리는 다음과 같이 구할 수 있다.

$$L_p^{norm}(A,B) \equiv \left(\sum_{i=1}^{n} |a_i - b_i - m|^p \right)^{1/p}$$

여기서 $m = \sum_{i=1}^{n} (a_i - b_i)/n$

시계열 데이타베이스에서 유사한 시퀀스를 검색하는데 있어 중요한 점은 검색 성능을 향상시키는 것이다. 일반적으로 시퀀스는 길이가 길고 고차원의 데이타이므로 시퀀스 검색에서 유사도 계산은 시간 소모적이고 많은 시스템 자원을 사용하게 된다. 데이타베이스의 크기가 커짐에 따라 순차 검색 방법은 성능이 매우 떨어지게 된다. 일반적으로 질의 시퀀스와 유사한 시퀀스의 빠른 검색을 위해 인덱싱이 사용된다. 다시 말해, 유사도 계산이 필요 없는 시퀀스를 제외하여 검색 공간을 신속히 줄이는 알고리즘이 필요하다. 일반적인 방식은 길이가 n인 시퀀스를 n-차원 공간 상의 한 점으로 사상한 후, R-tree[14] 또는 R*-tree[15]와 같은 공간 접근 기법을 이용하여 이러한 점을 인덱싱한다. 그러나 시퀀스와 같은 고차원 데이타를 그대로 공간 접근 기법을 이용하여 인덱싱하는 방식은 차원의 저주(dimensionality curse)[1,16]라는 현상에 의해 성능 저하를 유발한다. 다시 말해 대부분의 다차원 인덱스 구조들은 차원에 대해 지수적으로 검색 시간이 증가하며 차원이 매우 높은 경우 궁극적으로 순차 검색보다 성능이 떨어지게 된다.

이러한 문제를 해결하기 위해 여러 가지 차원 감소 기법들이 제안되어 왔다. 대표적인 차원 감소 기법으로는 DFT(Discrete Fourier Transform), DWT(Discrete Wavelet Transform, SVD(Singular Value Decomposition) 등이 있다. 이러한 차원 감소 기법에 의해 저차원으로 사상된 시퀀스들을 기존의 공간 접근 기법을 이용하여 인덱싱하는 방식이 일반적으로 시계열 데이타베이스의 검색에 주로 이용되고 있다. 차원 감소 기법을 이용한 시계열 데이타를 인덱싱할 때 가장 중요한 것은 순차 검색보다 효율적이면서 착오 누락(false dismissals)이 없다는 것을 보장해야 한다는 것이다[1,8-11,16-19]. 다시 말해 순차 검색을 할 때와 동일한 결과를 얻을 수 있어야 한다는 것을 의미한다. 착오 누락이 없다는 것을 보장하기 위해서는 추출된 특징 벡터 간의 거리가 실제 시퀀스 간의 거리보다 작거나 같아야 한다[1,8-11,16-19].

시퀀스 검색에서 다른 중요한 점은 여러 가지 Lp 거리 함수를 지원할 수 있어야 한다는 점이다[21]. DFT 또는 DWT와 같은 차원 감소 기법을 이용한 검색 방법들은 유클리드(L_2) 거리에서만 효과적인 것으로 알려져 있으며 다른 거리 함수에 대해서는 그 성능을 보장하지 못한다[1,16].

정규 거리에 기반한 유사 시퀀스 검색을 위해 기존의 기법들은 각 시퀀스의 평균을 이용하여 정규화를 시행한 후 정규화된 시퀀스에서 특징을 추출하여 인덱싱하는 방식을 채택하고 있다. 이러한 정규화 과정은 각 시퀀스의 평균을 구한 후 이를 시퀀스의 요소들에서 빼주는 과정을 거쳐야 한다.

본 논문에서는 정규화 과정 없이 정규 거리에 기반한 유사 시퀀스 검색 기법을 제안한다. 시퀀스에서 각 요소들의 상대적인 거리는 정규화 과정에서도 변하지 않으므로 이러한 성질을 이용하여 시퀀스에서 특징을 추출하여 R-tree와 같은 공간 접근 기법을 이용하여 인덱싱한다. 제안된 기법은 정규화 과정 없이 비슷한 형태의 시퀀스를 검색할 수 있으며 착오 누락이 없음을 보장한다.

본 논문의 구성은 다음과 같다. 2절에서 시퀀스 검색과 관련된 연구에 대해 살펴보고, 3절에서 제안된 기법에 대해 설명한다. 4절에서 제안된 기법의 성능을 평가

하며, 5절에서 결론을 맺는다.

2. 관련 연구

시계열 데이타의 빠른 검색과 매칭을 위해 다양한 기법들이 제안되어 왔다. 가장 인기 있는 방법은 고차원의 시퀀스를 차원 감소 기법을 사용하여 저차원으로 사상한 후 기존의 공간 접근 기법을 이용하여 인덱싱하는 방식이다.

시계열 데이타베이스에서 유사한 시퀀스를 검색하는 문제는 [1]에서 처음 제기되었다. 길이가 같은 시퀀스 데이타베이스에서 유사 검색 질의를 처리하기 위해 *F-index* 기법이 제안되었다. [1]에서는 DFT를 사용하여 차원 감소시킨 시퀀스를 R-tree를 사용하여 인덱싱하였다. [1]에서 제안된 시퀀스 검색 기법은 [8]에서 길이가 다른 시퀀스 검색을 위해 확장되었다. *ST-index*라 하는 이 기법은 일정 길이로 시퀀스를 슬라이딩 윈도우라 하는 부분 시퀀스들로 나눈 후 각 부분 시퀀스를 DFT로 차원 감소시킨 후 이를 인덱싱하였다. 두 논문 모두 유사도의 척도로는 유클리드 거리를 사용하였다. *ST-index* 기법의 성능을 향상시키기 위한 연구가 [12]에서 수행되었다. *duality*에 기반한 이 기법은 *ST-index*와 달리 질의 시퀀스를 슬라이딩 윈도우로 처리하여 성능을 개선시켰다.

[9]에서는 차원 감소 기법으로 DFT 대신 DWT를 사용하였다. DWT는 DFT에 비해 계산 복잡도가 간단하며 더 효과적인 특징 추출이 가능하다. 그러나 DWT는 길이가 2의 지수인 시퀀스에 대해서만 처리가 가능한 단점이 있다.

[20]에서는 *STB-index*라는 인덱싱 기법이 제안되었다. 이 기법에서는 시퀀스를 여러 개의 세그먼트로 나눈 후 각 세그먼트의 상태를 0 또는 1로 나타내었다. 만약 상승 상태이면 1로, 하강 상태면 0으로 나타낸 후 이를 인덱싱하였다. 이러한 인덱싱 기법은 유클리드 거리를 사용한 유사 검색과는 성격이 다르며 유클리드 거리 입장에서 보면 착오 누락이 발생하게 된다.

[21]에서는 SVD가 차원 감소 기법으로 사용되었다. SVD는 DFT나 DWT와 달리 전체 데이타 분포를 고려한 차원 감소 기법이기 때문에 효과적인 특징 추출이 가능하다. 그러나 계산 복잡도가 너무 높기 때문에 시간 공간적인 비용이 매우 큰 단점이 있다. 그리고 새로운 시퀀스가 데이타베이스에 추가되는 경우 인덱스 구성을 다시 해야 되는 단점이 있다.

[10, 11]에서는 시퀀스를 길이가 같은 여러 세그먼트로 나눈 후 각 세그먼트의 평균값을 특징으로 하는 차원 감소 기법이 제안되었다. 같은 개념의 독립적인 연구가 [16]에서 제안되었다.

[7, 17, 19, 22]에서는 유클리드 거리 대신 타임와핑(time warping) 거리를 유사도의 척도로 사용하였다. 타임와핑은 길이가 다른 시퀀스를 비교하는 데 있어 적합한 유사도의 척도를 제공한다. 그러나 타임와핑은 계산 복잡도가 높고, 삼각 부등식이 성립하지 않는 거리 함수이기 때문에 기존의 공간 접근 기법으로 인덱싱하는 것이 쉽지 않다. [7]에서는 suffix-tree를 사용하여 타임와핑에 기반한 시퀀스 검색을 지원하였다. 이는 suffix-tree가 어떠한 거리 함수도 가정하고 있지 않기 때문에 가능한 것으로 순차 검색에 비해 상당한 성능 향상을 보이고 있다. 그러나 suffix-tree를 구성하는 데 있어 원래 시퀀스 데이타의 크기보다 더 큰 인덱스 구조를 요구하는 단점이 있다.

기타 다양한 기법들이 효율적인 시퀀스 검색을 위하여 연구되었다[18,23-26].

3. 제안된 기법

본 논문에서는 정규 거리에 기반하여 시계열 데이타베이스에서 질의 시퀀스와 유사한 시퀀스를 찾는 문제에 초점을 두고 있다. 이 절에서는 수직 이동의 전처리 과정 없이 정규 거리에 기반한 특징 추출 기법을 소개하고, 제안된 기법이 착오 누락이 없음을 보이겠다.

3.1 특징 추출 기법

우리의 목표는 각 시퀀스에서 모양과 관련된 특징을 추출한 후 추출된 특징들 간의 거리가 정규 거리의 하한이 되도록 하는 것이다. 본 논문에서 제안하는 특징 추출 기법은 다음과 같이 길이가 n인 시퀀스를 길이가 l인 m개의 세그먼트로 나눈 후 각 세그먼트에서 변이의 평균을 특징으로 추출한다.

정의 1. [평균 변이 특징] 시퀀스 $A=(a_1, a_2, \cdots, a_n)$에서 각 세그먼트의 변이의 평균을 특징으로 추출하여 구성된 특징 벡터 FA는 다음과 같이 정의된다.

$$FA=\langle FA_1, FA_2, \cdots, FA_m\rangle$$

$$\equiv\langle \frac{1}{l-1}\cdot\sum_{i=1}^{(l-1)}|a_{i+1}-a_i|, \frac{1}{l-1}\cdot\sum_{i=l}^{2(l-1)}|a_{i+1}-a_i|, \cdots, \frac{1}{l-1}\cdot\sum_{i=(m-1)l+(2-m)}^{s(l-1)}|a_{i+1}-a_i|\rangle$$

제안된 특징 추출 기법은 다음의 관찰에 의한 것이다.

관찰 1. 세그먼트에서 변이의 평균은 정규화에 불변이다.

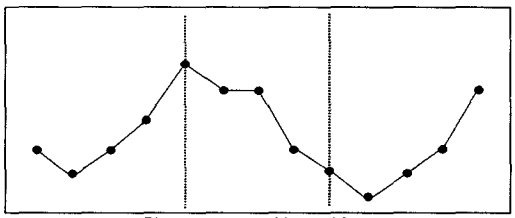
이에 대한 증명은 다음과 같다.

$$FA_m=\frac{1}{l-1}\cdot\sum_{i=(m-1)l+(2-m)}^{s(l-1)}|a_{i+1}-a_i|$$

$$=\frac{1}{l-1}\cdot\sum_{i=(m-1)l+(2-m)}^{s(l-1)}|(a_{i+1}-m_a)-(a_i-m_a)|$$

여기서 $m_a=\sum_{i=1}^{n}(a_i)/n$

그림 2는 본 논문에서 제안된 차원 감소를 위한 특징 추출 기법을 보여주고 있다. 길이가 13인 시퀀스가 3차원으로 사상되었다. 각 세그먼트에서 변이의 평균은 정규화에 대해 불변이므로 추출된 변이의 평균 특징을 이용하여 정규 거리에 기반한 질의를 처리할 수 있다.



FA = (평균변이(5 4 5 6 8), 평균변이(8 7 7 5 4), 평균변이(4 3 4 5 7))
= (1.25,1.1.25)

그림 2 제시된 특징 추출 기법

3.2 정규 거리의 하한

착오 누락이 없음을 보장하기 위해서는 특징 벡터 간의 거리 함수가 원래 시퀀스 간의 거리보다 작거나 같다는 것을 보여야 한다. 다음의 수식이 성립함을 보이는 것은 그다지 어렵지 않다.

$$L_p^{simple}(FA) \le L_p^{norm}(A)$$

그러나 추출된 특징들 간의 거리 함수가 느슨하므로 많은 착오 해답(false alarm)이 포함될 수 있다. 따라서 착오 해답을 효과적으로 제거하기 위해 특징 거리 함수를 하한에 밀착될 수 있는 방법을 찾아야 한다. 다시 말해 다음과 같은 요소 a_p를 찾아야 한다.

$$a_p \cdot L_p^{simple}(FA) \le L_p^{norm}(A), \quad a_p \ge 1$$

a_p는 Lp 거리 함수가 볼록 함수(convex function)라는 성질을 이용하여 구할 수 있다. 여기서 볼록 함수에 대한 수학적 결과를 이용한 정리를 [16,27]에서 가져와 이용한다.

정리 1. 실수 집합 R에 대해 $x_1, x_2, ..., x_n \in R$이며 $\lambda_1, \lambda_2, ..., \lambda_n \in R$, $\lambda_i \ge 0$, $(\sum_{i=1}^{n} \lambda_i) = 1$라고 하자. F()가 볼록 함수라면 다음의 부등식이 성립한다.

$F(\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_n x_n) \le \lambda_1 F(x_1) + \lambda_2 F(x_2) + ... + \lambda_n F(x_n)$

증명. [27]에 증명이 나와 있음.

정리 1을 이용하여 $\lambda_i = (1/n)$으로 두면, 다음의 추론(corollary)을 얻을 수 있다.

추론 1. 어떠한 시퀀스 $A=(a_1, a_2, \cdots, a_n)$, $1 \le p \le \infty$에 대해 다음이 성립한다.

$$\begin{aligned}
&(n-1) \cdot |\text{평균변이}(A)|^p \\
&\le \sum_{i=1}^{n-1} |a_{i+1} - a_i|^p \\
&\le 2^{p-1} \cdot \left(\sum_{i=1}^{n-1} |a_{i+1} - m_a|^p + \sum_{i=2}^{n} |a_{i+1} - m_a|^p \right) \\
&\le 2^p \cdot \sum_{i=1}^{n} |a_i - m_a|^p \\
&\text{여기서 } m_a = \sum_{i=1}^{n} (a_i)/n
\end{aligned}$$

또는 시퀀스 A의 길이가 l인 세그먼트 S_j에 다음이 성립한다.

$$\begin{aligned}
&(l-1) \cdot |\text{평균변이}(S_j)|^p \\
&\le \sum_{i=(j-1)l+(2-j)}^{j(l-1)} |a_{i+1} - a_i|^p \\
&\le 2^{p-1} \cdot \left(\sum_{i=(j-1)l+(2-j)}^{j(l-1)} |a_{i+1} - m_a|^p \right. \\
&\left. + \sum_{i=(j-1)l+(3-j)}^{j(l-1)+1} |a_{i+1} - m_a|^p \right) \\
&\text{여기서 } m_a = \sum_{i=1}^{n} (a_i)/n
\end{aligned}$$

위의 추론을 이용하여 본 논문의 주 정리를 이끌어 낼 수 있다.

정리 2. 어떠한 시퀀스 $A=(a_1, a_2, \cdots, a_n)$, $1 \le p \le \infty$에 대해 다음이 성립한다.

$$\begin{aligned}
&\sqrt[p]{n-1} \cdot L_p^{simple}(A) \\
&\le 2 \cdot \left(\sum_{i=1}^{n} |a_i - m_a|^p \right)^{1/p} \\
&= 2 \cdot L_p^{norm(A)} \\
&\text{여기서 } m_a = \sum_{i=1}^{n} (a_i)/n
\end{aligned}$$

증명 2. 추론 1을 이용하여 다음의 수식이 성립함을 보일 수 있다.

$$\begin{aligned}
&(l-1) \cdot L_p^{simple}(FA)^p = (l-1) \cdot \sum_{j=1}^{m} |\text{평균변이}(S_j)|^p \\
&\le 2^{p-1} \cdot \left(\sum_{i=1}^{l-1} |a_i - m_a|^p + \sum_{i=2}^{l} |a_i - m_a|^p \right) \\
&+ 2^{p-1} \cdot \left(\sum_{i=l}^{2(l-1)} |a_i - m_a|^p + \sum_{i=l+1}^{2(l-1)+1} |a_i - m_a|^p \right) \\
&+ \\
&\cdots\cdots \\
&+ 2^{p-1} \cdot \left(\sum_{i=(j-1)l+(2-j)}^{j(l-1)} |a_i - m_a|^p + \sum_{i=(j-1)l+(3-j)}^{j(l-1)+1} |a_i - m_a|^p \right) \\
&\le 2^p \cdot \sum_{1}^{n} |a_i - m_a|^p \\
&= 2^p \cdot L_p^{norm}(A)^p \\
&\text{여기서 } m_a = \sum_{i=1}^{n} (a_i)/n
\end{aligned}$$

정리 2의 결과에 따라 세그먼트의 평균 변이 특징을 이용하면 수직 이동의 전처리 과정 없이 정규 거리에 기반한 질의를 착오 누락 없이 처리할 수 있다. 다시 말해 두 시퀀스 A, B를 비교할 때, $L_p^{norm}(A,B) \le \varepsilon$의 질의는 $L_p^{simple}(FA,FB) \le 2 \cdot \varepsilon / \sqrt[p]{l-1}$와 동등한 질의이다. 따라서 $2/\sqrt[p]{l-1}$라는 요소를 이용하여 검색 공간을 더욱 효과적으로 축소할 수 있다. $a_p < 1$인 경우에는 사용될 수 없으며, 이 때는 기존의 특징 거리 함수를 사용한다.

3.3 질의 처리

이 절에서는 제안된 시퀀스 검색 기법의 전체 과정에 대해 설명한다. 질의가 수행되기 전에 먼저 각 시퀀스에서 특징이 추출되고, 추출된 특징을 이용하여 R-tree와

같은 다차원 인덱스 구조에 저장한다. 구성된 인덱스 구조를 이용하여 질의 시퀀스와 유사한 시퀀스의 후보들을 먼저 검색한 후 검색된 후보들에 대해서는 실제 거리를 계산하는 후처리 과정을 통해 정답을 출력한다. 알고리즘 1은 제안된 검색 기법을 기술하고 있다.

Algorithm 1. SMV-Indexing
Input : query sequence Q, error bound ε
Output : data sequences within error bound ε
Begin
result ← NULL; candidate ← NULL;
// project the query time sequence Q into the index space
FQ ← FeatureExtraction(Q);
// Candidate selection using a Spatial Access Method
candidate←candidate U IndexSearching(FQ, SAM, ε);
// Postprocessing to remove false alarms
// Ci ∈ candidate
for(i=1; i < size of candidate; i++)
if(ComputeNormDistance(Ci, Q) ≤ ε)
result ← Ci U result ;
else reject Ci ;
return result ;
End

질의 처리 과정은 두 단계로 구성되어 있다. 우선 인덱스 구조를 이용하여 질의 시퀀스와 유사할 가능성이 있는 후보 시퀀스들을 찾은 과정과 얻어진 후보들에 대해서 질의 시퀀스와의 실제 정규 거리를 계산하는 후처리 과정으로 구성되어 있다. 후처리 과정에서 실제 정규 거리가 주어진 오차 범위를 벗어나는 착오 해답을 제거하고 실제 정규 거리가 오차 범위 이내인 유사한 시퀀스들만을 사용자에게 질의 결과로 반환한다.

4. 성능 평가

이 절에서는 제안된 기법의 성능을 분석하기 위한 실험 결과를 제시한다. 제안된 기법의 성능을 평가하기 위해 순차 검색 기법과 검색 공간 비율과 질의 처리 시간 측면에서 비교해 보았다. 먼저 실험 환경에 대해 설명하고, 그 다음에 실험 결과를 제시하였다

4.1 실험 환경

제안 기법과 순차 검색 기법 모두 C++ 언어를 이용하여 500MHz dual Pentium III 중앙처리장치와 512M 주메모리, 40G 하드디스크를 가진 리눅스 기계에서 구현되었다. 다차원 인덱스 구조로는 Katatama의 R*-tree[1]를 이용하였다. 실험에 사용된 주식 데이타는 1998년 1월부터 2000년 3월까지의 내용을 가지고 있으며 서울 증권 시장에서 구한 것이다. 주식 데이타베이스는 길이가 128인 2000개의 일일종가의 시퀀스들로 구성되어 있다. 정규 거리에 기반하여 유사한 시퀀스 검색을 위해 25번의 질의를 수행하였으며, 질의 시퀀스는 데이타베이스에서 임의로 선택되었다.

4.2 실험 결과

먼저 실제 정규 거리 계산이 필요한 후보 시퀀스의 개수에 대해 제안 기법과 순차 검색 기법을 비교해 보았다. 인덱스 검색 과정에서 정답이 될 가능성이 없는 시퀀스를 제거하고 후보만을 걸러내는 여과 능력을 평가하기 위해 검색 공간 비율을 평가해보았다. 검색 공간 비율은 다음과 같이 정의된다.

$$\text{검색 공간 비율} = \frac{\text{후보 시퀀스의 개수}}{\text{전체 시퀀스의 개수}}$$

그림 3은 L_1, L_2, L_3 그리고 L_∞에서 오차 범위를 증가시키면서 검색 공간 비율을 측정해 본 결과를 나타내고 있다. 이 실험에서 추출된 특징의 차원은 8이 이용되었다. 그림 3의 실험 결과로 알 수 있듯이 제안된 기법이 효율적으로 검색 공간을 줄이고 있음을 보이고 있다.

그림 4는 L_1, L_2, L_3 그리고 L_∞에서 추출되는 특징의 차원을 변화시키면서 측정한 검색 공간 비율을 나타내고 있다. 이 실험에서 오차 범위는 5000으로 고정시켰다. 특징의 차원을 증가시킬수록 여과 능력이 증가됨을

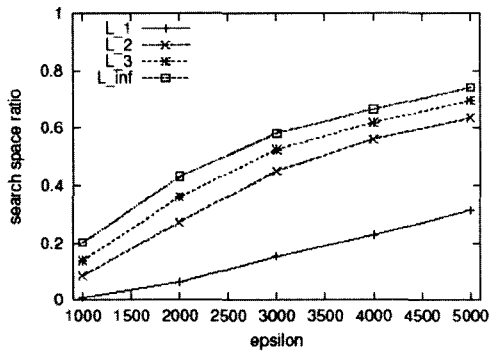


그림 3 오차 범위에 따른 검색 공간 비율

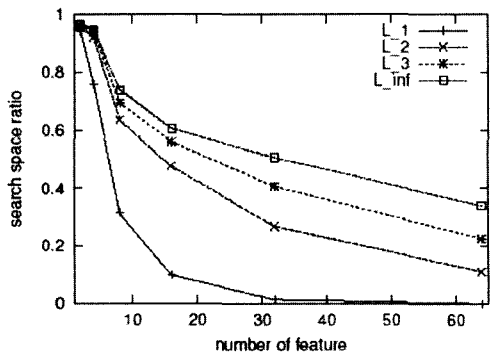


그림 4 특징 차원에 따른 검색 공간 비율

1) Katatama의 R*-tree 소스 코드는 다음의 사이트에서 구할 수 있다. http://research.nii.ac.jp/~katayama/homepage/research/srtree/English.html

알 수 있다.

특징 공간에서 사용되는 차원의 수가 많을수록 여과 능력은 향상된다. 그러나 이러한 특징을 실제 공간 접근 기법을 사용하여 인덱싱하는 경우, 여과 능력이 크다고 하더라도 항상 좋은 성능을 보장하는 것은 아니다. 이는 다차원 인덱스 구조는 차원의 증가에 따라 급속한 성능 저하를 보이게 되는 차원의 저주가 발생하기 때문이다. 이러한 차원의 저주 문제를 검증하기 위해 L_2에서 8차원의 특징을 사용한 제안 기법과 16차원의 특징을 사용한 제안 기법, 그리고 순차 검색 기법을 검색 공간 비율과 평균 질의 처리 시간의 측면에서 비교해 보았다.

그림 5와 그림 6은 16차원의 특징을 사용한 제안 기법이 8차원의 특징을 사용한 경우보다 여과 능력이 더 크지만, 차원의 저주로 인해 실제 질의 처리 시간은 별 다른 차이가 없음을 보여주고 있다.

오차 범위가 커짐에 따라 질의 처리 시간이 증가하는데, 이것은 오차 범위가 커짐에 따라 후보로 선택되는 시퀀스의 수가 증가하게 되어 후처리 과정에서 후보 시퀀스들 중에서 착오 해답을 제거하는 시간이 증가하기 때문이다.

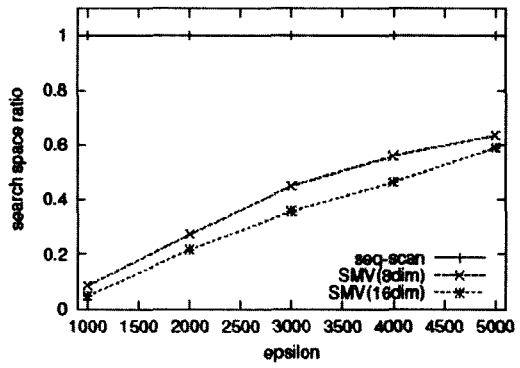


그림 5 L_2에서의 검색 공간 비율

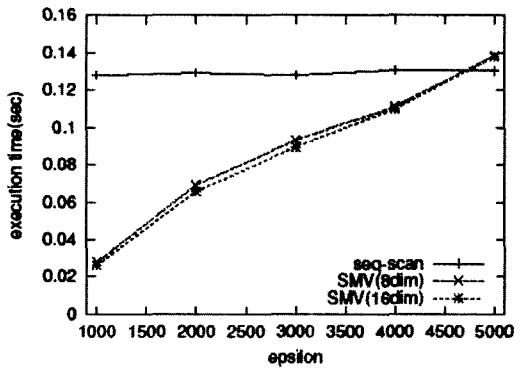


그림 6 L_2에서의 질의 처리 시간

5. 결 론

단순한 Lp 거리는 시퀀스 검색에 있어 유사도로 주로 사용되어 왔으나 시퀀스의 절대값에 의해 영향을 쉽게 받는 단점이 있다. 다시 말해, 비슷한 모양의 시퀀스라 할 지라도 수직적인 위치가 다른 경우 유사하지 않다고 분류하는 단점이 있다. 단순한 Lp 거리의 단점을 해결하기 위해 정규 거리가 제안되어 왔다. 그러나 정규 거리에 기반한 기존의 유사도 검색 기법은 평균을 이용한 정규화 과정을 필요로 하고 있다. 본 논문에서는 정규화 과정 없이 임의 Lp 정규 거리에 기반한 유사 검색 기법을 제안하였다. 시퀀스의 세그먼트 요소들의 상대적인 변이는 정규화에 불변이라는 속성을 이용하여 각 세그먼트로부터 변이의 평균을 특징으로 추출하였다. 추출된 특징 벡터를 이용하여 거리 함수를 정의하고 이를 이용하여 효율적인 유사 검색을 지원하였다.

참 고 문 헌

[1] Rakesh Agrawal, Christos Faloutsos, Arun N. Swami, "Efficient Similarity Search In Sequence Databases," In Proceedings of FODO, pp. 69~84, 1993.

[2] Rakesh Agrawal, T. Imielinski, Arun N. Swami, "Database Mining: A Performance Perspective," IEEE TKDE, Special issue on Learning and Discovery in Knowledge-Based Databases 5(6), pp. 914~925, 1993.

[3] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," In Proceedings of KDD conference, pp. 82~88, 1996.

[4] Davood Rafiei, Alberto O. Mendelzon, "Similarity-Based Queries for Time Series Data," In Proceedings of ACM SIGMOD Conference, pp. 12~25, 1997.

[5] Kelvin Kam Wing Chu, Sze Kin Lam, Man Hon Wong, "An Efficient Hash-Based Algorithm for Sequence Data Searching," The Computer Journal 41(6), pp. 402~415, 1998.

[6] Davood Rafiei, "On Similarity-Based Queries for Time Series Data," In Proceedings of ICDE, pp. 410~417, 1999.

[7] Sanghyun Park, Wesley W. Chu, Jeehee Yoon, Chihcheng Hsu, "Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases," In Proceedings of ICDE pp. 23~32, 2000.

[8] Christos Faloutsos, M. Ranganathan, Yannis. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases," In Proceedings of ACM SIGMOD Conference, pp. 419~429, 1994.

[9] Kin-pong Chan, Ada Wai-chee Fu, "Efficient

Time Series Matching by Wavelets," In Proceedings of ICDE 1999: 126~133.
[10] Eamonn J. Keogh, Michael J. Pazzani, "A Simple Dimensionality Reduction Technique for Fast Similarity Search in Large Time Series Databases," In Proceedings of PAKDD Conference, pp. 122~133, 2000.
[11] Eamonn J. Keogh, Kaushik Chakrabarti, Sharad Mehrotra, Michael J. Pazzani, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases," In Proceedings of ACM SIGMOD Conference, pp. 151~162, 2001.
[12] Yang-Sae Moon, Kyu-Young Whang, Woong-Kee Loh, "Duality-Based Subsequence Matching in Time-Series Databases," In Proceedings of ICDE, pp. 263~272, 2001.
[13] Sze Kin Lam, Man Hon Wong, "A Fast Projection Algorithm for Sequence Data Searching," DKE 28(3), pp. 321~339, 1998.
[14] Antonin Guttman, "R-trees: A Dynamic Index Structure for Spatial Searching," In Proceedings of ACM SIGMOD Conference, pp. 47~57, 1984.
[15] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, Bernhard Seeger, "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles," In Proceedings of ACM SIGMOD Conference, pp. 322~331, 1990.
[16] Byoung-Kee Yi, Christos Faloutsos, "Fast Time Sequence Indexing for Arbitrary Lp Norms," In Proceedings of VLDB Conference, pp. 385~394, 2000.
[17] Sangwook Kim, Sanghyun Park and W. Chu, "An Index-based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," In Proceedings of ICDE, pp. 607~614, 2001.
[18] Sangjun Lee, Dongseop Kwon, Sukho Lee, "Efficient Similarity Search for Time Series Data Based on the Minimum Distanc," In Proceedings of CAiSE, pp. 377~391, 2002.
[19] Eamonn J. Keogh, "Exact Indexing of Dyanmic Time Warping," In Proceedings of VLDB Conference, pp. 406~417, 2002.
[20] Eamonn J. Keogh, Michael J. Pazzani, "An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback," In Proceedings of KDD Conference, pp. 239~243, 1998.
[21] Flip Korn, H. V. Jagadish, Christos Faloutsos, "Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences," In Proceedings of ACM SIGMOD Conference, pp. 289~300, 1997.
[22] Byoung-Kee Yi, H. V. Jagadish, Christos Faloutsos, "Efficient Retrieval of Similar Time Sequences Under Time Warping," In Proceedings of ICDE, pp. 201~208, 1998.
[23] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, Kyuseok Shim, "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," In Proceedings of VLDB Conference, pp. 490~501, 1995.
[24] Chung-Sheng Li, Philip S. Yu, Vittorio Castelli, "HierarchyScan: A Hierarchical Similarity Search Algorithm for Databases of Long Sequences," In Proceedings of ICDE, pp. 546~553, 1996.
[25] Kelvin Kam Wing Chu, Man Hon Wong, "Fast Time-Series Searching with Scaling and Shifting," In Proceedings of PODS, pp. 237~248, 1999.
[26] Chang-Shing Perng, Haixun Wang, Sylvia R. Zhang, D. Stott Parker, "Landmarks: a New Model for Similarity-based Pattern Querying in Time Series Databases," In Proceedings of ICDE, pp. 33~42, 2000.
[27] M. H. Protter, C. B. Morrey, "A First Course in Real Analysis," Springer-Verlag, 1997.

이 상 준

1996년 서울대학교 컴퓨터공학과 졸업(공학사). 1998년 서울대학교 대학원 컴퓨터공학과 졸업(공학석사). 1998년~현재 서울대학교 대학원 전기컴퓨터공학부 박사과정. 관심분야는 멀티미디어 시스템, 이동객체 및 시계열 데이타베이스

이 석 호

1964년 연세대학교 정치외교학과 졸업 1975년, 1979년 미국 텍사스대학교 전산학 석사와 박사학위 취득. 1979년~1982년 한국과학원 전산학과 조교수. 1982년~1986년 한국정보과학회 논문 편집위원장. 1986년~1988년 한국정보과학회 부회장. 1988년~1989년 미국 IBM T.J. Watson연구소 객원교수. 1988년~1990년 데이타베이스연구회 운영위원장 1989년~1991년 서울대학교 중앙교육연구전산원 원장. 1994년 한국정보과학회 회장. 1997년~현재 한국학술진흥재단 부설 첨단학술정보센터 소장. 1982년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이타베이스, 멀티미디어 데이타베이스