한국 주식 데이터를 이용한 서브시퀀스 매칭 방법의 효과성 평가

유 승 근[†] · 이 상 호[††]

요 약

기존의 서브시퀀스 매칭 방법은 검색을 효율적으로 수행하기 위한 인덱스 구성 방법에 대하여 연구하였으며, 서브시퀀스 매칭 방법의 효과성 평가를 고려하지 않았다. 본 논문은 서브시퀀스 매칭 방법의 효과성에 대하여 고려하였으며, 서브시퀀스 매칭 방법의 효과성을 평가 할 수 있는 2가지 척도를 제안한다. 한국 주식 데이터와 5가지 서브시퀀스 매칭 방법에 대하여 제안된 효과성 측정 방안을 적용하였으며, 그 결과를 분석하였다. 실험 결과, 정규화를 지원하는 서브시퀀스 매칭 방법과 스케일링과 쉬프팅 변환을 지원하는 서브시퀀스 매칭 방법이 상대적으로 효과적인 서브시퀀스를 검색하였다.

키워드 : 데이터 마이닝, 데이터 시퀀스, 시계열 데이터, 효과성 평가

Effectiveness Evaluations of Subsequence Matching Methods Using KOSPI Data

Seung Keun Yoo[†] · Sang Ho Lee[††]

ABSTRACT

Previous researches on subsequence matching have been focused on how to make indexes in order to speed up the matching time, and do not take into account the effectiveness issues of subsequence matching methods. This paper considers the effectiveness of subsequence matching methods and proposes two metrics for effectiveness evaluations of subsequence matching algorithms. We have applied the proposed metrics to Korean stock data and five known matching algorithms. The analysis on the empirical data shows that two methods (i.e., the method supporting normalization, and the method supporting scaling and shifting) outperform the others in terms of the effectiveness of subsequence matching.

Key Words : Data Mining, Data Sequence, Time-series Data, Effectiveness Evaluation

1. 서 론

시계열 데이터(time-series data)는 일정한 시간 주기마다 얻어진 실수 값들로 이루어진 데이터이며, 주가 및 환율 데이터, 날씨 정보 데이터, 제품 판매 데이터, 의료 측정 데이터 등의 예가 있다[9, 12]. 컴퓨터의 계산 및 저장능력이 발전함에 따라 많은 양의 시계열 데이터를 활용하고자 하는 연구가 활발하게 이루어져 왔으며, 특히 시계열 데이터에 대한 유사 시퀀스 매칭은 데이터 마이닝의 중요한 분야로 자리 잡고 있다[1, 4, 5, 9].

시계열 데이터베이스에 저장된 시계열 데이터를 데이터 시퀀스(data sequence)라 부르며, 유사 시퀀스 매칭(similar sequence matching)은 사용자가 입력한 유사 허용치(ε) 이하로 질의 시퀀스(query sequence)와 유사한 데이터 시퀀스를 시계열 데이터베이스에서 검색하는 방법이다[1, 4]. 유사 시퀀스 매칭은 전체 매칭(whole matching)과 서브시퀀스 매칭(subsequence matching) 알고리즘으로 구분된다[4]. 전체 매칭 알고리즘은 질의 시퀀스와 동일한 길이를 가지는 여러 개의 데이터 시퀀스 중에서 질의 시퀀스와의 거리가 사용자가 입력한 유사 허용치보다 작거나 같은 시퀀스를 검색하는 알고리즘이다. 서브시퀀스 매칭 알고리즘은 서로 다른 길이를 가지는 여러 개의 데이터 시퀀스 중에서 질의 시퀀스와의 거리가 사용자가 입력한 유사 허용치보다 작거나 같은 서브시퀀스를 검색하는 알고리즘이다. 일반적으로, 서브시퀀스 매칭은 전체 매칭에 비하여 보다 다양한 분야에서 응용될 수 있다[13].

[13]에서는 서브시퀀스 매칭 알고리즘을 전처리(preprocessing) 변환에 따라 구분하였다. [1, 4]에서는 데이터 시퀀스와 질의 시퀀스간의 전처리 변환 없이 비교하는 알고리즘을 제안하였다. [2, 3, 8, 12, 13]에서는 데이터 시퀀스와 질의 시퀀스에 대하여 스케일링(scaling), 쉬프팅(shifting), 정규화(normalization), 타임 워핑(time warping) 등의 변환을 수행한 결과를 비교하는 알고리즘을 제안하였다. 이러한 전처리 변환의 목적은 응용 분야에 부합되도록 시퀀스 간의 유사성

※ 본 연구는 숭실대학교 교내 연구비 지원으로 이루어졌습니다.
† 준 회 원 : 숭실대학교 대학원 컴퓨터학과
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수
논문접수 : 2005년 1월 11일, 심사완료 : 2005년 3월 24일

정의에 유연성을 주는 것이다[13].

기존의 서브시퀀스 매칭을 다룬 논문에서는 슬라이딩 윈도우를 이용하여 서브시퀀스를 생성하였으며, 검색 조건하에서 유사 서브시퀀스를 검색하는 경우 시작 위치가 서로 연속적인 트리비얼 매치(trivial match)를 유사 서브시퀀스로 판정하였다. 서브시퀀스의 길이가 w인 경우, 연속적인 2개의 트리비얼 매치는 w-1개의 엔트리들이 서로 중복되기 때문에 적절한 방법을 사용하여 처리되어야만 한다. 또한 다차원 인덱스(multidimensional index) 구성을 통한 효율적인 서브시퀀스 검색 방법들을 제안하였으나, 실험에 사용한 서브시퀀스 매칭 방법의 효과성을 평가하지 않았다.

본 논문에서는 유사 서브시퀀스의 새로운 정의를 통하여 트리비얼 매치를 처리하고, 서브시퀀스 매칭 방법의 효과성을 평가하기 위한 새로운 평가 척도를 제안한다. 한국 주식 데이터에 기존의 5가지 서브시퀀스 매칭 방법을 사용하여 서브시퀀스를 검색한 후, 검색된 서브시퀀스에 2가지 척도를 적용한 실험 결과를 분석한다. 5가지 서브시퀀스 매칭 방법은 정규화 변환, 타임 워핑 변환, 스케일링 및 쉬프팅 변환을 지원하는 유사 서브시퀀스 매칭 방법과 코사인 유사도(cosine similarity)를 이용하는 유사 서브시퀀스 매칭 방법 및 순차 검색(sequential scan)이다. 또한 사용자가 입력하는 유사 허용치를 질의 시퀀스의 길이에 따른 백분율로 사용함으로써 사용자에게 유사 허용치 입력의 편리성을 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 5가지 서브시퀀스 매칭 방법에 대하여 기술하며, 3장에서는 새로운 유사 서브시퀀스를 정의하고, 유사 허용치를 질의 시퀀스의 길이에 따라 백분율로 사용하는 방법을 기술한다. 서브시퀀스 매칭 방법의 효과성 평가를 위한 2가지 평가 척도를 제안한다. 검색된 서브시퀀스에 평가 척도를 적용한 결과를 보이며, 이를 분석한다. 4장에서는 본 연구에 대한 결론을 맺는다.

2. 서브시퀀스 매칭 방법

유사 서브시퀀스를 검색하기 위해서는 데이터 시퀀스를 동일한 길이의 서브시퀀스로 나누어야한다. 데이터 시퀀스를 $\vec{S}$, 데이터 시퀀스의 길이를 n, 데이터 시퀀스 $\vec{S}$의 k번째 엔트리를 $\vec{S}[k](1 \leq k \leq n)$라 하면, 데이터 시퀀스 $\vec{S}$를 크기 w인 슬라이딩 윈도우로 나눈다 함은 $1 \leq i \leq n$인 모든 i에 대하여 $\vec{S}[i]$를 시작 위치로 하는 윈도우들을 구성함을 의미한다[9]. 이렇게 구성된 윈도우들을 서브시퀀스라 정의한다. <표 1>은 본 논문에서 사용되는 표기법(notation)을 정리한 것이다.

■ 순차 검색

순차 검색은 질의 시퀀스와 유사한 서브시퀀스를 검색하기 위한 가장 쉬운 방법이다. 데이터 길이 n($1 \leq n$)인 서브시퀀스 $\vec{S}=\{s_1, s_2, ..., s_n\}$와 질의 시퀀스 $\vec{Q}=\{q_1, q_2, ..., q_n\}$는 n차원 공간상의 한 점으로 표현이 가능함으로, 식 (1)을 사용하여 비교하는 두 시퀀스간의 유클리디안 거리(euclidean distance)를 계산한다.

<표 1> 주요 표기법 정리

기 호	정 의
n	데이터 시퀀스의 길이
w	슬라이딩 윈도우의 길이
$\vec{Q}$	질의 시퀀스
$\vec{S}_i$	i번째 서브시퀀스($1 \leq i \leq n-w+1$)
$Len(\vec{S}_i)$	서브시퀀스 $\vec{S}_i$의 길이
$\vec{S}_i[k]$	서브시퀀스의 k번째 엔트리($1 \leq k \leq Len(\vec{S}_i)$)
$D_{similar}(\vec{Q})$	질의 시퀀스 $\vec{Q}$와 유사한 서브시퀀스의 집합

$$D=\sqrt{\sum_{i=1}^{n}(s_i-q_i)^2} \quad (1)$$

데이터 시퀀스 $\vec{S}$를 질의 시퀀스와 동일한 길이인 슬라이딩 윈도우로 나누고, 생성된 모든 서브시퀀스와 질의 시퀀스에 대하여 유클리디안 거리를 계산한다. 비교하는 두 시퀀스간의 유클리디안 거리가 사용자가 입력한 유사 허용치보다 작거나 같으면, 두 시퀀스는 서로 유사하다고 판정한다.

■ 스케일링과 쉬프팅 변환을 지원하는 서브시퀀스 매칭 방법

[3]에서는 스케일링과 쉬프팅 변환을 지원하는 유사 서브시퀀스 매칭 방법을 제안하였다. 스케일링 변환은 시퀀스를 구성하는 모든 엔트리에 동일한 실수 값을 곱하여 시퀀스의 진동폭을 변화시키는 변환이다. 쉬프팅 변환은 시퀀스를 구성하는 모든 엔트리에 동일한 실수 값을 증감하여 시퀀스를 수직으로 이동시키는 변환이다. 데이터 시퀀스 $\vec{S}$를 질의 시퀀스와 동일한 길이인 슬라이딩 윈도우로 나누고, 질의 시퀀스와 모든 서브시퀀스에 대하여 SE(shift-eliminated)변환을 수행한다. 식 (2)는 시퀀스 $\vec{S}$의 SE변환식이다.

$$T_{se}=\vec{S}-\frac{\vec{S}\cdot\vec{N}}{\|\vec{N}\|^2}\vec{N} \quad (2)$$

SE변환은 질의 시퀀스와 서브시퀀스를 SE평면상의 직선과 점으로 변환하므로, 직선과 점 사이의 최소 거리를 계산한다. 계산된 거리가 사용자가 입력한 유사 허용치보다 작거나 같으면, 두 시퀀스는 유사하다고 판정한다. 본 논문에서는 스케일링과 쉬프팅 변환을 지원하는 서브시퀀스 매칭 방법을 "SEMatch"라 한다.

■ 정규화 변환을 지원하는 서브시퀀스 매칭 방법

[8, 13]에서는 정규화 변환을 지원하는 유사 서브시퀀스

매칭 방법을 제안하였다. 정규화 변환은 시퀀스를 구성하는 값들이 변화하는 패턴을 비교하는 데에 유용하다[8]. 데이터 시퀀스 $\vec{S}$를 질의 시퀀스와 동일한 길이인 슬라이딩 윈도우로 나누고, 생성된 모든 서브시퀀스와 질의 시퀀스에 대하여 정규화 변환을 수행한다. 길이 n인 서브시퀀스 $\vec{S}$를 정규화 변환한 서브시퀀스 $v(\vec{S}) = (\tilde{S}[i])$는 다음의 식 (3)과같이 정의한다[8, 13].

$$\tilde{S}[i] = \frac{\vec{S}[i] - \mu(\vec{S})}{\delta(\vec{S})} \tag{3}$$

정규화 변환된 모든 서브시퀀스 $v(\vec{S}_i)$와 $v(\vec{Q})$간의 유클리디안 거리를 계산한다. 계산된 거리가 사용자가 입력한 유사 허용치 보다 작거나 같으면, 두 시퀀스는 유사하다고 판정한다. 본 논문에서는 정규화 변환을 지원하는 서브시퀀스 매칭 방법을 "NormalMatch"라 한다.

■ 타임 워핑 변환을 지원하는 서브시퀀스 매칭 방법

[2][12]에서는 타임 워핑 변환을 지원하는 유사 서브시퀀스 매칭 방법을 제안하였다. 타임 워핑 변환은 시퀀스 내의 각 엔트리 값을 임의의 수만큼 반복시키는 것을 허용하는 변환이다. 두 시퀀스 $\vec{S}$와 $\vec{Q}$간의 타임 워핑 거리는 식 (4)와 같이 순환식(recurrence relation) $r_{tw}(X, Y)$ $(X = 1, 2, \ldots, |\vec{S}|, Y = 1, 2, \ldots, |\vec{Q}|)$을 기반으로 한 동적 프로그래밍(dynamic programming)을 이용하여 효율적으로 계산될 수 있다[2].

$$r_{tw}(x,y) = D_{base}(\vec{S}[x], \vec{Q}[y]) + \min\begin{cases} r_{tw}(x, y-1), \\ r_{tw}(x-1, y), \\ r_{tw}(x-1, y-1), \end{cases} \tag{4}$$

데이터 시퀀스 $\vec{S}$를 질의 시퀀스와 동일한 길이인 슬라이딩 윈도우로 나누고, 생성된 모든 서브시퀀스와 질의 시퀀스에 대하여 타임 워핑 거리를 계산한다. 동적 타임 워핑 거리가 사용자가 입력한 유사 허용치 보다 작거나 같으면, 두 시퀀스는 유사하다고 판정한다. 본 논문에서는 동적 프로그래밍을 이용하여 타임 워핑 변환을 지원하는 서브시퀀스 매칭 방법을 "DTWMatch"라 한다.

■ 코사인 유사도를 이용한 서브시퀀스 매칭 방법

코사인 유사도(cosine similarity)는 벡터 이론(vector theory)을 바탕으로 한 기법으로서, 정보 검색 분야에서 서로 다른 두 문서 또는 문서와 질의 간의 유사도를 측정하는데 사용한다. 코사인 유사도에 사용되는 벡터 이론은 두 벡터의 내적(inner product)과 벡터의 길이(length)이다. 두 벡터 $\vec{V}$와 $\vec{W}$의 내적 $\vec{V} \cdot \vec{W}$과 길이 $\|\vec{V}\|$의 정의는 식 (5)와 같다.

$$\vec{V} \cdot \vec{W} = \sqrt{\sum_{i=1}^{n} (\vec{V}[i] \times \vec{W}[i])^2}, \quad \|\vec{V}\| = \sqrt{\sum_{i=1}^{n} (\vec{V}[i])^2} \tag{5}$$

두 벡터가 이루는 각의 코사인 유사도는 식 (6)과 같이 두 벡터의 내적을 두 벡터의 길이의 곱으로 나눈 값이다. 코사인 유사도를 이용해서 유사 서브시퀀스를 검색하는 방법을 "CMatch"라 한다.

$$\cos\theta = \frac{\vec{S} \cdot \vec{Q}}{\|\vec{S}\| \times \|\vec{Q}\|} \tag{6}$$

"CMatch"의 유사 서브시퀀스 검색 방법은 다음과 같다. 데이터 시퀀스 $\vec{S}$를 질의 시퀀스와 동일한 길이인 슬라이딩 윈도우로 나누고, 생성된 모든 서브시퀀스와 질의 시퀀스에 대하여 두 시퀀스가 이루는 각의 코사인 값을 계산한다. 비교하는 두 시퀀스 $\vec{S}_i$와 $\vec{Q}$가 이루는 각의 코사인 값이 사용자가 입력한 유사 허용치 값보다 작거나 같으면 두 시퀀스는 유사하다고 판정한다.

3. 유사 서브시퀀스 정의 및 매칭 방법의 효과성 평가

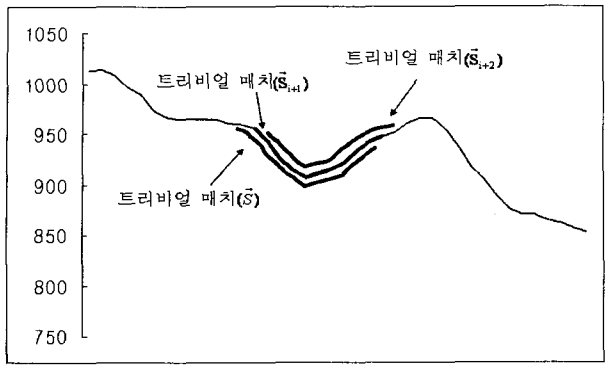
3.1 유사 서브시퀀스의 정의

기존의 서브시퀀스 매칭 방법들은 전체 데이터 시퀀스에서 서브시퀀스를 생성하기 위해서 질의 시퀀스와 동일한 길이의 슬라이딩 윈도우로 전체 데이터 시퀀스를 나눈다. 서브시퀀스 $\vec{S}_i$와 질의 시퀀스 $\vec{Q}$간의 유클리디안 거리 혹은 두 시퀀스가 이루는 각의 코사인 값을 계산하거나, 전처리 과정을 거친 시퀀스 $t(\vec{S}_i)$와 $t(\vec{Q})$간의 유클리디안 거리, 타임 워핑 거리, 점과 직선의 거리를 계산하여 사용자가 입력한 유사 허용치보다 작거나 같으면, 두 시퀀스 $\vec{S}_i$와 $\vec{Q}$는 유사하다고 판정한다. 이와 같은 유사 서브시퀀스 매칭 방법은 슬라이딩 윈도우를 이용하여 서브시퀀스를 생성하고 유사 허용치 이하로 범위(range) 검색을 수행하기 때문에, 트리비얼 매치가 발생하는 문제가 있다.

정의 1. 질의 시퀀스 $\vec{Q}$에 대하여 검색 조건을 만족하는 연속된 서브시퀀스 $\vec{S}_i$, $\vec{S}_{i+1}$, ..., $\vec{S}_{i+j}$ $(j \geq 1)$가 존재하면, 트리비얼 매치가 발생한다고 한다. ■

예제 1. (그림 1)은 질의 시퀀스 $\vec{Q}$에 대하여 검색 조건을 만족하는 3개의 서브시퀀스가 연속적으로 발생하는 트리비얼 매치의 예를 보인다.

트리비얼 매치는 한 개의 질의 시퀀스에 대하여 검색 조건을 만족하는 서브시퀀스 검색 시 여러 번 발생할 수 있다.

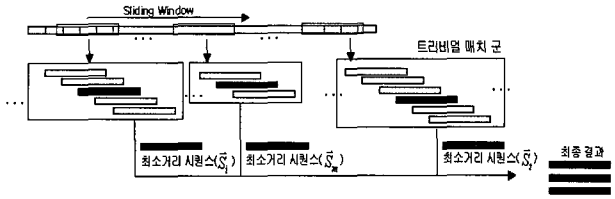


(그림 1) 트리비얼 매치의 예

변화의 폭이 적은 시계열 데이터는 변화의 폭이 큰 시계열 데이터에 비해 트리비얼 매치가 많이 발생하며, 시계열 데이터의 노이즈가 심한 경우는 노이즈가 적은 경우에 비해 적은 양의 트리비얼 매치가 발생한다[6]. 트리비얼 매치는 슬라이딩 윈도우 이용에 의한 적합하지 않은 매칭 결과이므로, 본 논문에서는 트리비얼 매치를 적절한 방법을 사용하여 처리한다.

정의 2. 트리비얼 매치가 발생하면, 트리비얼 매치를 발생시키는 연속된 서브시퀀스 중에서 질의 시퀀스와의 유사도 값이 최소인 한 개의 서브시퀀스만을 유사 서브시퀀스로 판정한다. ■

예제 2. (그림 2)는 3개의 트리비얼 매치 군이 발생하였을 경우 유사 서브시퀀스 정의를 통해 트리비얼 매치를 처리하는 과정이다. 슬라이딩 윈도우를 사용하여 서브시퀀스를 생성한 후, 질의 시퀀스와의 거리가 유사 허용치보다 작은 서브시퀀스를 검색한다.

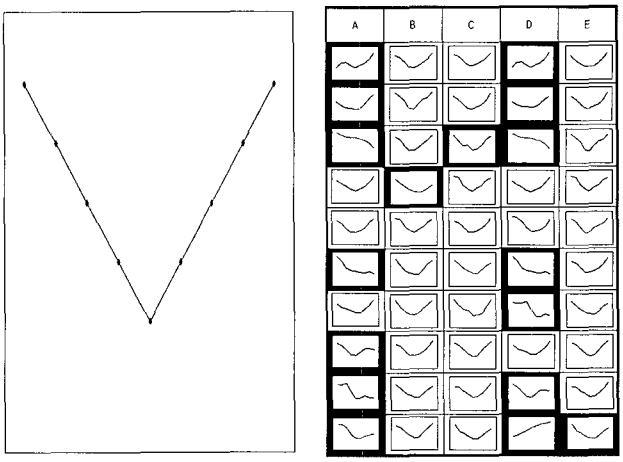


(그림 2) 유사 서브 시퀀스 판정의 예

검색 결과 3개의 트리비얼 매치 군이 발생하였으며, 각 군에서 질의 시퀀스와의 거리가 최소인 서브시퀀스 S_i, S_m, S_l를 유사 서브시퀀스로 판정한다. ■

3.2 서브시퀀스 매칭 방법의 효과성 평가 척도

(그림 3)의 (a)는 V자 모양의 질의 시퀀스이며, (b)는 A, B, C, D, E의 5가지 서브시퀀스 매칭 방법에 V자 모양의 질의 시퀀스를 입력하였을 경우 검색된 10개의 유사 서브시퀀스의 모양을 도식화한 것이다. (그림 3)의 (b)에서 음영처리된 셀(cell)은 직관적으로 판단할 때 질의 시퀀스의 모양과 유사하지 않은 서브시퀀스의 모양을 나타낸다. 음영처리된 서브시퀀스가 검색된 결과에 적게 포함될수록 효과적인 서브시퀀스 매칭 방법이라고 말할 수 있다. (그림 3)의 (b)와 같이 검색된 서브시퀀스의 모양을 도식화하여 질의 시퀀스와 비교를 수행하는 직관적인 방법은 평가자에 의존적이며, 많은 시간을 필요로 하는 작업이기 때문에 본 논문에서는 서브시퀀스 매칭 방법을 평가할 수 있는 2가지의 평가 척도를 제안한다.

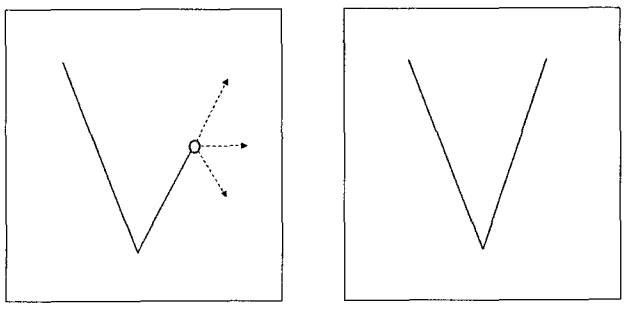


(a) V자 모양의 질의 시퀀스 (b) 검색된 서브 시퀀스의 모양

(그림 3) V자 모양의 질의 시퀀스 및 검색된 서브시퀀스의 모양

본 논문에서 제안하는 유사 서브시퀀스 매칭 방법의 평가 척도는 "포함 관계 평가 척도"와 "추세 예측 평가 척도"이다. "포함 관계 평가 척도"는 질의 시퀀스를 사용하여 서브시퀀스를 검색하고, 동일한 유사 허용치내에서 질의 시퀀스 길이의 70%, 80%, 90%로 서브 질의 시퀀스를 생성하여 검색된 결과를 비교한다. (그림 4)의 (a)는 V자형 질의 시퀀스 길이의 80%에 해당하는 서브 질의 시퀀스의 모양 및 이후 추세를 나타낸다.

서브 질의 시퀀스 이후의 추세는 (그림 4)의 (a)에서와 같이 상승, 보합, 하락으로 구분 될 수 있으며, 서브 질의 시



(a) 80% 서브 질의 시퀀스 (b) 질의 시퀀스

(그림 4) 포함 관계 평가 척도에서 포함율의 의미

퀀스를 사용하여 검색된 결과는 3가지 추세에 해당하는 서브시퀀스를 포함한다. 포함율은 질의 시퀀스를 사용하여 검색된 결과 중에서 몇 개의 서브시퀀스가 서브 질의 시퀀스를 사용하여 검색된 결과에 포함되는지를 분석하는 것이다. "포함 관계 평가 척도"의 단점은 서로 모양이 다른 두개의 질의 시퀀스 $\vec{Q}_1$, $\vec{Q}_2$가 있을 때, 이 두 질의 시퀀스의 서브시퀀스는 $a\vec{Q}_1$, $b\vec{Q}_2(a,b \leq 1)$의 모양은 서로 같을 수 있다는 것이다.

정의 3. "포함 관계 평가 척도"에서 포함율의 정의는 식 (7)과 같다.

$$\text{포함율} = \frac{count(S_{similar}(\vec{Q}) \cap S_{similar}(a\vec{Q}))}{count(S_{similar}(\vec{Q}))} \ (a \leq 1) \quad (7)$$

식 (7)에서 분모는 질의 시퀀스를 사용하여 검색된 서브시퀀스의 개수이고, 분자는 질의 시퀀스를 사용하여 검색된 서브시퀀스와 서브 질의 시퀀스를 사용하여 검색된 서브시퀀스 중에서 공통적으로 검색된 서브시퀀스의 개수이다. ■

예제 3. "포함관계 평가 척도"에서의 포함율 계산 예이다. 질의 시퀀스를 사용하여 검색된 서브시퀀스의 개수가 10개이고, 질의 시퀀스길이의 80%인 질의 시퀀스를 사용하여 검색된 서브시퀀스의 개수가 15개이다. 두 집합에서 동시에 검색된 개수가 6개라고 한다면,

$count(S_{similar}(\vec{Q})) = 10$,

$count(S_{similar}(\vec{Q}) \cap S_{similar}(a\vec{Q})) = 6$

이므로, 포함율은 60%이다. ■

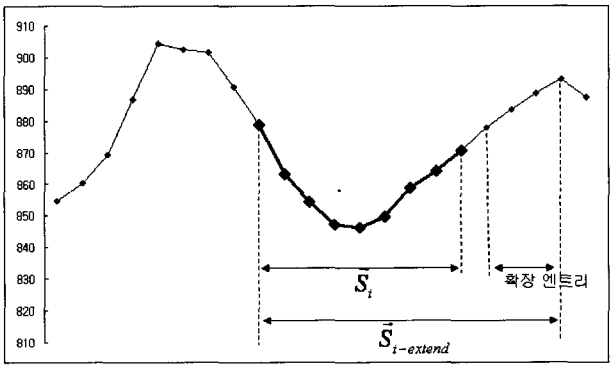
"추세 예측 평가 척도"는 질의 시퀀스가 암시하는 추세와 검색된 서브시퀀스 이후의 추세가 일치하는가를 판단하고, 추세가 일치한다면 일치하는 기간을 이용하여 예측율을 계산한다. 예측율은 서브시퀀스마다 계산되므로 예측율의 평균을 사용하여 매칭 방법의 효과성을 평가한다. 만약 검색된 서브시퀀스의 추세가 질의 시퀀스의 추세를 따르지 않는다면, 식 (9)에서 확장 서브시퀀스의 길이와 서브시퀀스의 길이가 동일하기 때문에 예측율은 1이다.

정의 4. "추세 예측 평가 척도"에서 예측율의 정의는 식 (8)과 같다.

$$\text{예측율} = \frac{Len(\vec{S}_{i-extend})}{Len(\vec{S}_i)} \quad (8)$$

단, $Len(\vec{S}_{i-extend})$는 검색된 서브시퀀스 이후의 추세가 질의 시퀀스와 암시하는 추세와 같을 때, 추세를 따르는 엔트리들의 개수와 검색된 서브시퀀스의 개수를 합산한 값이다. ■

예제 4. (그림 5)는 검색된 서브시퀀스 $\vec{S}_i$와 확장 서브시퀀스 $\vec{S}_{i-extend}$를 나타내며, "추세 예측 평가 척도"에서의 예측율 계산 예이다.



(그림 5) 서브시퀀스와 확장 서브시퀀스의 예

V자형 질의 시퀀스를 사용하여 서브시퀀스를 검색한 결과, 유사 서브시퀀스로 $\vec{S}_i$가 검색되었다. V자형 질의 시퀀스가 상승 추세를 암시하므로, 검색된 서브시퀀스 이후의 상승 추세 엔트리들은 (그림 5)에서의 확장 엔트리이다. 확장 엔트리의 길이는 4이므로, 검색된 서브시퀀스의 예측율은 1.44 (= (9+4) / 9)이다. ■

예측율의 평균은 확장 엔트리의 개수가 많은 몇 개의 유사 서브시퀀스에 영향을 받으므로, "추세 예측 평가 척도"에서는 "추세 예측 실패율"을 고려한다.

정의 5. "추세 예측 실패율" 정의는 식 (9)와 같다.

$$\text{추세예측실패율} = \frac{count(\vec{S}_{p=1}(\vec{Q}))}{count(\vec{S}_{similar}(\vec{Q}))} \quad (9)$$

단, $count(\vec{S}_{p=1}(\vec{Q}))$는 질의 시퀀스 $\vec{Q}$와 유사한 서브시퀀스의 집합에서 예측율이 1인 유사 서브시퀀스의 개수이다. ■

예제 5. "추세 예측 실패율"의 계산 예이다. 질의 시퀀스를 사용하여 검색된 서브시퀀스의 개수가 10개이고, 이중에서 예측율이 1인 서브시퀀스의 개수가 4개라면,

$count(S_{similar}(\vec{Q})) = 10$,

$count(S_{p=1}(\vec{Q})) = 4$

이므로, "추세 예측 실패율"은 40%이다. ■

3.3 사용자 입력 유사 허용치 계산

각 서브시퀀스 매칭 방법은 서브시퀀스와 질의 시퀀스 간

의 유사도 측정 방법과 전처리 변환 과정의 수행 여부에 따라 각기 다른 범위의 유사도 측정값이 계산된다. 질의 시퀀스의 길이가 증가하면, 그에 따른 유사도 계산 값도 커지므로, 유사 서브시퀀스 검색을 수행하기 위해서 사용자는 각 서브시퀀스 매칭 방법에서 계산된 유사도 값의 범위에 해당하는 적절한 유사 허용치 값을 입력하여야 한다. 이러한 문제는 질의 시퀀스의 길이에 따른 백분율로 유사 허용치를 제공함으로써 해결 될 수 있다. 순차 검색 및 "DTWMatch", "SEMatch"에서는 식 (10)과 같이 질의 시퀀스 크기에 대한 백분율로 ε'를 계산한다.

$$\varepsilon' = \|\vec{Q}\| \cdot \frac{\epsilon}{100} \tag{10}$$

정규화 변환을 지원하는 "NormalMatch"에서는 시퀀스의 평균과 표준 편차를 이용하여 시퀀스를 정규화 변환한다. 벡터 공간에서의 정규화 변환식의 기하학적 의미는 벡터 $\vec{V}$를 $\mu(\vec{V})$만큼 수직 이동 시킨 후, 벡터의 크기를 $\sigma(\vec{V})$로 나눈 것이다. 수직 이동한 벡터의 길이는 변하지 않으므로, "NormalMatch"에서의 ε'는 식 (11)과 같이 계산한다.

$$\varepsilon' = \frac{\|\vec{Q}\|}{\sigma(\vec{Q})} \cdot \frac{\epsilon}{100} \tag{11}$$

"CMatch"에서는 서브시퀀스 $\vec{S_i}$와 질의 시퀀스 $\vec{Q}$간의 유사도 계산 값 D의 값이 작기 때문에, 질의 시퀀스 $\vec{Q}$를 정규화 변환한 시퀀스, $v(\vec{Q})$를 이용하여 식 (12)와 같이 ε'를 계산한다.

$$\varepsilon' = \|v(\vec{Q})\| \cdot \frac{\epsilon}{100} \tag{12}$$

각 유사 서브시퀀스의 매칭 방법의 특징은 <표 2>와 같다.

3.4 실험 데이터 및 질의 시퀀스

실험에 사용된 주식 데이터는 1997년 1월부터 2003년 12월까지의 일별 종합 주가 지수(Korea Composite Stock Price Index, KOSPI)이다. 잡음을 제거하기 위하여 5일 이동평균이 적용된 종가(closing price)를 사용하여 시계열 데이터베이스를 구축하였으며, 1997년부터 2003년 12월까지의 거래일수는 1811일이다. 본 논문에서는 일별 종합 주가 지수를 주식 데이터라 부른다.

질의 시퀀스는 실제 주식 데이터가 아닌 합성(synthetic) 데이터이다. 합성 데이터를 질의 시퀀스로 사용하는 이유는 실제 주식 데이터의 일부로 질의 시퀀스를 사용하는 경우 결과에 항상 질의 시퀀스가 포함되어지며, 이는 사용자가 이미 알고 있는 서브시퀀스로서 분석이 필요 없다. 또한 질의 시퀀스의 잡음을 제거하기 위한 작업이 필요하며, 사용자가 사용하고자 하는 질의 시퀀스의 모양을 데이터 시퀀스 내에서 찾기가 어렵다는 단점이 있기 때문이다.

실험에서 질의 시퀀스로 사용한 패턴은 일반적으로 주식 데이터의 기술적 분석에서 많이 사용되어 지는 패턴으로, 길이 9일 및 15일의 V자형 패턴, 길이 12일의 하락 N자형 패턴, 길이 19일의 이중 바닥형 패턴이다. 질의 시퀀스를 구성하는 엔트리들의 범위는 한국 주식 데이터의 중간값인 710과 740사이로 하였다.

(그림 6)의 (a)는 9일 길이의 V자형 질의 시퀀스이며, (b)는 15일 길이의 V자형 질의 시퀀스이다. 9일 길이 V자형 시퀀스는 5일 하락 추세와 5일 상승 추세로 구성되며, 생성 방법은 다음과 같다.

(1) 주식 데이터에 대하여 5일 길이의 슬라이딩 윈도우를 사용하여 서브시퀀스를 생성한다.
(2) 모든 서브시퀀스의 첫 번째 엔트리와 마지막 엔트리의 차를 서브시퀀스의 변동값이라고 한다.
(3) 변동값의 평균을 동일한 추세를 구성하는 엔트리간의 변동 구간의 개수로 나눈다.
(4) 과정 (3)에서 계산된 값을 동일한 추세를 구성하는 서브시퀀스 엔트리간의 변동값으로 하여 질의 시퀀스를 생성한다.

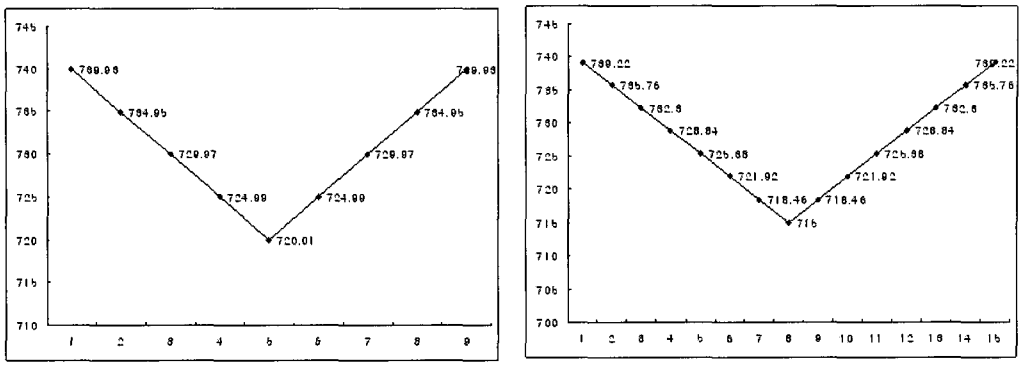
15일 길이 V자형 시퀀스는 8일 하락 추세와 8일 상승 추세로 구성되며, 이후의 생성 과정은 9일 길이의 V자형 시퀀스와 같다.

(그림 7)은 12일 길이의 하락 N자형 질의 시퀀스의 모양 및 엔트리 값이다. 12일 길이의 하락 N자형 질의 시퀀스의 경우는 5일의 하락 추세와 4일의 상승 추세 그리고 5일의 하락 추세로 구성되며, 생성 방법은 다음과 같다.

(1) 주식 데이터에 대하여 5일 길이의 슬라이딩 윈도우를

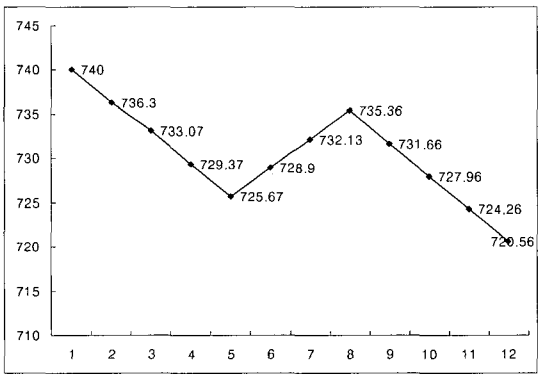
〈표 2〉 유사 서브시퀀스 매칭 방법의 특징 요약

	Sequential Scan	NormalMatch	SEMatch	TWMatch	CMatch
Scaling	X	O	O	X	O
Shifting	X	O	O	X	X
길이관계	동일길이	동일길이	동일길이	임의의 길이	동일길이
임계 값	$\|\vec{Q}\| \cdot \frac{\epsilon}{100}$	$\frac{\|\vec{Q}\|}{\sigma(\vec{Q})} \cdot \frac{\epsilon}{100}$	$\|\vec{Q}\| \cdot \frac{\epsilon}{100}$	$\|\vec{Q}\| \cdot \frac{\epsilon}{100}$	$\|v(\vec{Q})\| \cdot \frac{\epsilon}{100}$
거리 계산	유클리디안 거리	유클리디안 거리	점과 직선의 거리	타임 워핑 거리	코사인 값

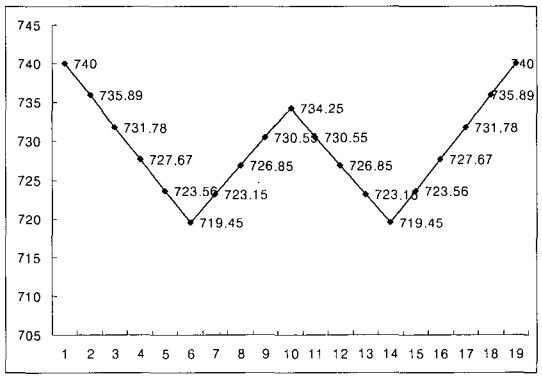


(a) 9일 길이의 V자형 시퀀스 (b) 15일 길이의 V자형 시퀀스

(그림 6) V자형 질의 시퀀스



(그림 7) 하락 N자형 질의 시퀀스



(그림 8) 이중 바닥형 질의 시퀀스

사용하여 서브시퀀스를 생성한다.

(2) 모든 서브시퀀스의 첫 번째 엔트리와 마지막 엔트리의 차를 서브시퀀스의 변동값이라고 한다.

(3) 변동값의 평균을 동일한 추세를 구성하는 엔트리간의 변동 구간의 개수로 나눈다.

(4) 과정 (3)에서 계산된 값을 동일한 추세를 구성하는 서브시퀀스 엔트리간의 변동값으로 하여 질의 시퀀스를 생성한다.

(그림 8)은 19일 길이의 이중 바닥형 질의 시퀀스의 모양 및 엔트리 값이다. 19일 길이의 이중 바닥형 질의 시퀀스 경우는 6일의 하락 추세, 5일의 상승 추세, 5일의 하락 추세, 6일의 상승 추세로 구성되며, 생성 방법은 다음과 같다.

(1) 주식 데이터에 대하여 5일 길이의 슬라이딩 윈도우를 사용하여 서브시퀀스를 생성한다.

(2) 모든 서브시퀀스의 첫 번째 엔트리와 마지막 엔트리의 차를 서브시퀀스의 변동값이라고 한다.

(3) 변동값의 평균을 동일한 추세를 구성하는 엔트리간의 변동 구간의 개수로 나눈다.

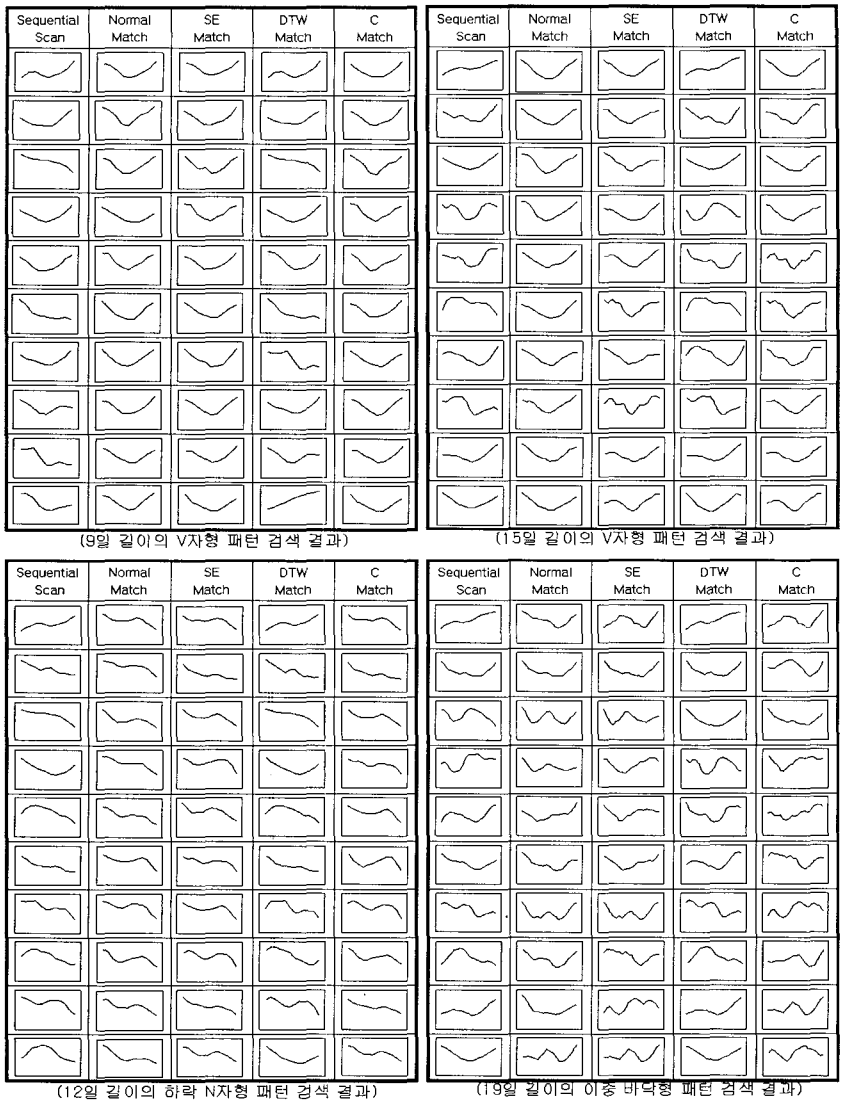
(4) 과정 (3)에서 계산된 값을 동일한 추세를 구성하는 서브시퀀스 엔트리간의 변동값으로 하여 질의 시퀀스를 생성한다.

3.5 실험 및 결과 분석

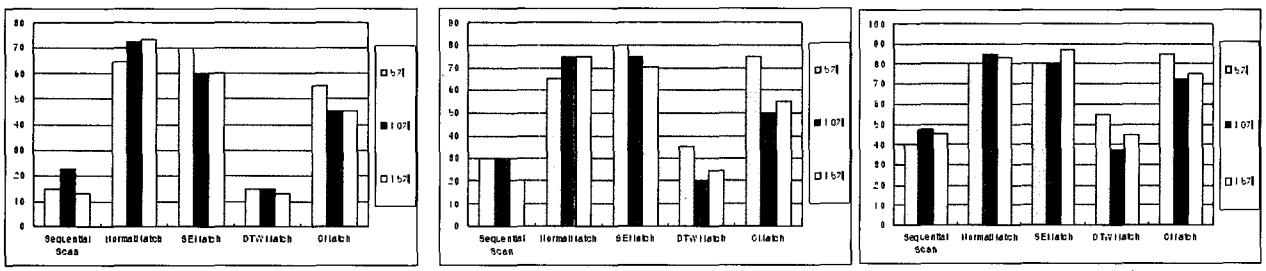
유사 서브시퀀스 매칭 방법에서 유사 허용치 값을 증가시키면서 검색된 5개, 10개, 15개의 유사 서브시퀀스를 실험 대상으로 하였다. 이중 바닥형 질의 시퀀스를 사용하여 유사 서브시퀀스를 검색하는 경우 순차 검색과 "DTWMatch"에서 검색된 최대 서브시퀀스의 개수가 12개이므로, 12개의 서브시퀀스를 실험 대상으로 하였다. (그림 9)는 각 서브시퀀스 매칭 방법에서 검색된 서브시퀀스의 모양을 시퀀스의 첫 번째 엔트리에 해당하는 날짜 별로 정렬하여 도식화한 것이다.

(그림 10)의 (a), (b), (c)는 70%, 80%, 90%의 서브 질의 시퀀스를 사용하여 "포함 관계 평가 척도"를 적용한 실험 결과이다. X축은 서브시퀀스 매칭 방법을 나타내고, Y축은 각 서브시퀀스 매칭 방법에 해당하는 평균 포함율이다. 서브 질의 시퀀스의 길이가 길어질수록 그래프의 세로축에 해당하는 평균 포함율도 증가하였다. 검색된 서브시퀀스의 개수가 5개이고, 70% 서브 질의 시퀀스와 80% 서브 질의 시퀀스를 사용한 실험 결과에서 "CMatch"의 포함율이 가장 높았으나, 나머지 경우는 항상 "NormalMatch"와 "SEMatch"의 평균 포함율이 가장 높았다. 순차 검색과 "DTWMatch"의 경우는 "NormalMatch"와 "SEMath"에 비해 상당히 낮은 평균 포함율 값을 보였다.

"NormalMatch"와 "SEMatch"가 상대적으로 높은 평균



(그림 9) 검색된 서브시퀀스의 모양

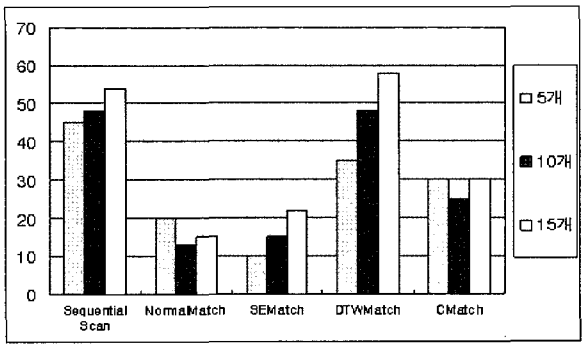


(a) 70% 서브 질의 시퀀스 실험 결과 (b) 80% 서브 실의 시퀀스 실험 결과 (c) 90% 서브 질의 시퀀스 실험 결과

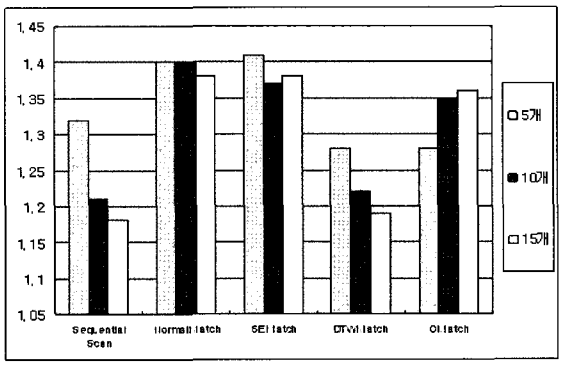
(그림 10) "포함 관계 평가 척도"의 적용 결과 그래프

포함율을 보인 것은 두 방법 모두 정규화 변환과 SE변환의 전처리 과정을 수행하므로 시퀀스를 구성하는 엔트리간 값 차이가 줄어, 계산되는 유사도 값이 전처리 과정을 거치지 않는 다른 방법에 비해 질의 시퀀스의 길이에 영향을 덜 받기 때문이라고 생각된다.

(그림 11)의 (a)는 각 서브시퀀스 매칭 방법의 "추세 예측 실패율"로 X축은 각 서브시퀀스 매칭 방법이며, Y축은 각 서브시퀀스 매칭 방법에서의 "추세 예측 실패율"의 평균을 나타낸다. (그림 11)의 (b)는 "추세 예측 평가 척도"를 적용한 실험의 결과 그래프로 X축은 서브시퀀스 매칭 방법이며, Y축은 각 서브시퀀스 매칭 방법에 해당하는 평균 추세 예측율이다.



(a) 추세 예측 실패율



(b) 추세 예측 평가 척도의 결과

(그림 11) "추세 예측 실패율"과 "추세 예측 평가 척도"의 적용 결과 그래프

(그림 11)의 (a)에서 "NormalMatch"와 "SEMatch"는 검색된 서브시퀀스 개수에 관계없이 25%미만의 추세 예측 실패율을 보였으며, 순차 검색과 "DTWMath", "CMatch"는 25%이상의 예측 실패율을 보였다. (그림 11)의 (b)에서 "NormalMatch"와 "SEMatch"가 상대적으로 질이 시퀀스가 암시하는 추세를 잘 예측하였다. 이는 "NormalMatch"와 "SEMatch"에서 검색된 서브시퀀스에 해당하는 확장 서브시퀀스의 길이가 길어서가 아니며, 예측된 추세를 따르지 않는 서브시퀀스가 결과 집합에 적게 포함되기 때문이다. "NormalMatch"와 "SEMatch"에서 검색된 서브시퀀스가 질의 시퀀스가 암시하는 추세를 잘 따르는 이유는 (그림 9)에서와 같이 검색된 서브시퀀스의 모양이 다른 서브시퀀스 매칭 방법에서 검색된 서브시퀀스의 모양보다 질의 시퀀스의 모양과 더 유사하기 때문이다.

각 서브시퀀스 매칭 방법에서 검색되어진 서브시퀀스에 "포함 관계 평가 척도"와 "추세 예측 평가 척도"를 적용한 결과 "NormalMatch"와 "SEMatch"가 상대적으로 평균 추세 예측율과 평균 포함율이 높았으며, 질의 시퀀스의 모양과 유사한 모양을 잘 찾아주었다.

4. 결 론

기존의 서브시퀀스 매칭을 다룬 논문에서는 트리비얼 매치를 유사 서브시퀀스로 판정하였으며, 인덱스 구성을 통하여 빠른 검색 속도를 제공하기위한 효율적인 서브시퀀스 검색 방법들을 제안하였으나 검색된 서브시퀀스에 대한 평가를 고려하지 않았다.

본 논문에서는 유사 서브시퀀스의 새로운 정의를 통하여 트리비얼 매치를 처리하는 방법을 보였으며, 유사 서브시퀀스 매칭 방법의 효과성을 평가할 수 있는 2가지 평가 척도를 제안하였다. 1997년부터 2003년까지의 한국 주식 데이터와 기존의 서브시퀀스 매칭 방법인 순차 검색, "NormalMatch", "SEMatch", "DTWMatch", "CMatch"를 사용하여, 질의 시퀀스와 유사한 서브시퀀스를 검색하였다. 본 논문에서 제안한 2가지의 서브시퀀스 매칭 방법의 효과성 평가 척도를 검색된 유사 서브시퀀스에 적용하고, 그 결과를 분석하였다. 실험 결과 "NormalMatch"와 "SEMatch"가 상대적으로 효과적인 서브시퀀스를 검색하였다. 또한 질의 시퀀스 길이의 백분율을 사용하여 유사 허용치를 입력함으로써 사용자에게 유사 허용치 입력의 편리성을 제공하였다.

참 고 문 헌

[1] R. Agrawal, C. Faloutsos and A. Swami, "Efficient Similarity Search in Sequence Databases," In Proceedings of the International Conference on Foundations of Data Organization and Algorithms, pp.69-84, 1993.

[2] D. J. Bernidt and J. Clifford, "Finding Patterns in Time Series : A Dynamic Programming Approach," Advances in Knowledge Discovery and Data mining, AAA/MIT Press, pp.229-248, 1996.

[3] K. K. W. Chu and M. H. Wong, "Fast Time-Series Searching with Scaling and Shifting," In Proceeding of the International Symposium on Principles of Databases Systems, pp.237-248, 1999.

[4] C. Faloutsos, M. Ranganathan and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Databases," In Proceedings of the International Conference on Management of Data, pp.419-429, 1994.

[5] D. Q. Goldin and P. C. Kanellakis, "On Similarity Queries for Time-Series Data : Constraint Specification and Implementation," In Proceedings of the International Conference on Principles of Data Mining and Knowledge Discovery, pp. 88-100, 1997.

[6] E. Keogh, J. Lin and W. Tuppel, "Clustering of Time Series Subsequence is Meaningless : Implication for Previous and Future Research," In Proceedings of the third IEEE International Conference on Data Mining, pp.115-125, 2003.

[7] R. R. Korfhage, "Information Storage and Retrieval," Wiley Press, 1997.

[8] W. K. Loh, S. W. Kim and K. Y. Whang, "Index Interpolation : An Approach for Subsequence Matching Supporting

Normalization Transform in Time-Series Databases," In Proceedings of the International Conference on Information and Knowledge Management, pp.480-487, 2000.

[9] Y. S. Moon, K. Y. Whang and W. K. Loh, "Duality-Based Subsequence Matching in Time-Series Databases," In Proceedings of the International Conference on Data Engineering, pp.263-272, 2001.

[10] S. H. Park, W. W. Chu, J. H. Yoon and C. Hsu, "Efficient Searches for Similar Subsequence of Different Lengths in Sequence Databases," In Proceedings of the International Conference on Data Engineering, pp.23-32, 2000.

[11] D. Rafiei and A. Mendelzon, "Similarity-Based Queries for Time Series Data," In Proceedings of the International Conference on Management of Data, pp.13-24, 1997.

[12] 김상욱, 박상현, "시퀀스 데이터베이스에서 타임 워핑을 지원하는 효과적인 유사 검색 기법", 정보과학회 논문지, 제28권 제4호, pp.643-654, 2001.

[13] 노웅기, 김상욱, 황규영, "시계열 데이터베이스에서 인덱스 보간법을 기반으로 정규화변환을 지원하는 서브시퀀스 매칭 알고리즘", 정보과학회 논문지, 제28권 제2호, pp.217-232, 2001.

유 승 근

e-mail : takeroots@hanmail.net

2003년 숭실대학교 컴퓨터학부(학사)

2005년 숭실대학교 대학원 컴퓨터학과 (석사)

관심분야 : 시계열 데이터 마이닝, 유사 서브시퀀스 매칭

이 상 호

e-mail : shlee@comp.ssu.ac.kr

1984년 서울대학교 전산공학과(학사)

1986년 미국 노스웨스턴대 전산학과 (석사)

1989년 미국 노스웨스턴대 전산학과 (박사)

1990년~1992년 한국전자통신연구원, 선임연구원

1999년~2000년 미국 조지메이슨대, 소프트웨어정보공학과, 교환 교수

1992년~현재 숭실대학교 컴퓨터학부 교수

관심분야 : 인터넷 데이터베이스, 데이터베이스 시스템 성능 평가 및 튜닝