시계열 데이터베이스에서 서브시퀀스 매칭을 위한 후처리 과정의 최적화

김 상 욱†

요 약

서브시퀀스 매칭은 주어진 질의 시퀀스와 변화의 추세가 유사한 서브시퀀스들을 시계열 데이터베이스로부터 검색하는 연산이며, 인덱스 검색 과정과 후처리 과정으로 구성된다. 본 논문에서는 서브시퀀스 매칭을 위한 후처리 과정의 최적화 방안에 관하여 논의한다. 기존의 서브시퀀스 매칭 기법들의 후처리 과정에서 발생하는 공통적인 문제점은 인덱스 검색 과정에서 각 후보 서브시퀀스가 반환될 때마다 이들이 최종 결과에 포함되는가에 대한 여부를 판별하기 위하여 질의 시퀀스와 비교한다는 것이다. 이러한 처리 방식은 후보 서브시퀀스들을 포함하는 동일한 시퀀스를 디스크로부터 여러 번 액세스되도록 할 뿐만 아니라 동일한 후보 서브시퀀스를 질의 시퀀스와 여러 번 비교하도록 한다. 따라서 이러한 중복 작업은 서브시퀀스 매칭의 처리 성능을 심각하게 저하시키는 중요한 원인이 된다. 본 연구에서는 이러한 문제점을 해결하는 새로운 최적의 기법을 제안한다. 제안된 기법은 인덱스 검색 과정에서 반환되는 모든 후보 서브시퀀스들을 이진 탐색 트리 내에 저장하고, 인덱스 검색 과정이 완료된 후에 일괄 처리 방식으로 후처리 작업을 수행한다. 이와 같은 일괄 처리 방식을 채택함으로써 제안된 기법은 위에서 언급한 중복 작업을 완전히 제거할 수 있다. 제안된 기법의 성능 개선 효과를 검증하기 위하여 실제 주식 데이터를 위한 다양한 실험을 수행한다. 실험 결과에 의하면, 제안된 기법은 기존의 기법과 비교하여 55배에서 156배까지의 성능 개선 효과가 있는 것으로 나타났다.

Optimization of Post-Processing for Subsequence Matching in Time-Series Databases

Sang-Wook Kim†

ABSTRACT

Subsequence matching, which consists of index searching and post-processing steps, is an operation that finds those subsequences whose changing patterns are similar to that of a given query sequence from a time-series database. This paper discusses optimization of post-processing for subsequence matching. The common problem occurred in post-processing of previous methods is to compare the candidate subsequence with the query sequence for discarding false alarms whenever each candidate subsequence appears during index searching. This makes a sequence containing candidate subsequences to be accessed multiple times from disk, and also have a candidate subsequence to be compared with the query sequence multiple times. These redundancies cause the performance of subsequence matching to degrade seriously. In this paper, we propose a new optimal method for resolving the problem. The proposed method stores all the candidate subsequences returned by index searching into a binary search tree, and performs post-processing in a batch fashion after finishing the index searching. By this method, we are able to completely eliminate the redundancies mentioned above. For verifying the performance improvement effect of the proposed method, we perform extensive experiments using a real-life stock data set. The results reveal that the proposed method achieves 55 times to 156 times speedup over the previous methods.

키워드 : 시계열 데이터베이스(Time-Series Databases), 서브시퀀스 매칭(Subsequence Matching), 후처리 과정(Post-Processing)

1. 서 론

시계열 데이터베이스(time-series databases)란 각 객체의 변화되는 값들의 연속으로 구성된 데이터 시퀀스(data sequences)들의 집합이며, 유사 검색(similarity search)이란 주어진 질의 시퀀스(query sequence)와 변화의 추세가 유사한 시퀀스들을 검색하는 연산이다[1, 4, 10]. 유사 검색은 시계열 데이터베이스를 기반으로 하는 데이터 마이닝(data mining) 및 데이터 웨어하우징(data warehousing) 분야에서 중요한 연산으로 사용된다[3, 6, 7, 10].

※ 본 연구는 학술진흥재단 선도연구자 지원사업(과제번호 : KRF-2000-041-E00258)과 한국과학재단 목적기초 연구지원사업(과제번호 : R05-2002-000-01085-0)의 연구비 지원으로 수행되었습니다.
† 정 회 원 : 강원대학교 컴퓨터정보통신공학부 부교수
논문접수 : 2001년 11월 17일, 심사완료 : 2002년 4월 9일

유사 검색은 크게 전체 매칭(whole matching)과 서브시퀀스 매칭(subsequence matching)으로 분류된다[1, 4, 9]. 전체 매칭은 모든 데이터 시퀀스들과 질의 시퀀스의 길이가 항상 동일하다는 전제하에 질의 시퀀스와 유사한 시퀀스를 데이터베이스로부터 검색한다. 반면, 부분 매칭은 다양한 길이의 데이터 시퀀스들이 존재하는 것을 허용하며, 데이터베이스로부터 질의 시퀀스와 유사한 서브시퀀스를 포함하는 시퀀스와 그 시퀀스 내에서의 유사 서브시퀀스의 시작 오프셋을 검색한다. 이와 같이, 서브시퀀스 매칭은 길이에 대한 제약이 없으므로 다양한 실제 응용 분야에서 널리 사용된다.

서브시퀀스 매칭은 인덱스 검색 과정과 후처리 과정으로 구성된다. 본 연구에서는 서브시퀀스 매칭을 위한 후처리 과정의 최적화 방안에 관하여 논의한다. 기존의 서브시퀀스 매칭 기법들의 후처리 과정에서 발생하는 공통적인 문제점은 인덱스 검색 과정에서 각 후보 서브시퀀스가 반환될 때마다 이들이 최종 결과에 포함되는가에 대한 여부를 판별하기 위하여 질의 시퀀스와 비교한다는 것이다. 이러한 처리 방식은 후보 서브시퀀스들을 포함하는 동일한 시퀀스를 디스크로부터 여러 번 액세스되도록 할뿐만 아니라 동일한 후보 서브시퀀스를 질의 시퀀스와 여러 번 비교하도록 한다. 따라서 이러한 중복 작업은 서브시퀀스 매칭의 처리 성능을 심각하게 저하시키는 중요한 원인이 된다.

본 연구에서는 이러한 문제점을 해결하는 새로운 최적의 기법을 제안한다. 제안된 기법은 인덱스 검색 과정에서 반환되는 모든 후보 서브시퀀스들을 이진 탐색 트리 내에 저장하고, 인덱스 검색 과정이 완료된 후에 일괄 처리방식으로 후처리 작업을 수행한다. 이와 같은 일괄 처리방식을 채택함으로써 제안된 기법은 위에서 언급한 중복 작업을 완전히 제거할 수 있다. 제안된 기법의 성능개선 효과를 검증하기 위하여 실제 주식 데이터를 위한 다양한 실험을 수행한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구로서 기존의 서브시퀀스 매칭 기법에 대하여 간략히 요약한다. 제 3장에서는 본 연구의 동기로서 기존 기법의 후처리 과정에서 나타나는 공통적인 성능상의 문제점을 지적한다. 제 4장에서는 이러한 문제점을 근본적으로 해결할 수 있는 새로운 기법을 제시한다. 제 5장에서는 제안된 기법의 성능 개선 효과를 보이기 위한 성능 평가 결과를 제시한다. 제 6장에서는 본 논문을 요약하고, 결론을 내린다.

2. 관련 연구

본 장에서는 본 연구에서 해결하고자 하는 문제를 공식적으로 정의하고, 이와 관련된 기존의 연구 결과를 요약한다.

2.1 문제 정의

시계열 데이터베이스 D, 길이 n의 질의 시퀀스 Q = (q[0], q[1], ···, q[n-1]), 그리고, 유사 허용치 ε이 주어질 때, *서브시퀀스 매칭 문제*는 다음과 같이 정의된다. D에 저장된 길이 N의 임의의 시퀀스 S = (s[0], s[1], ···, s[N-1]) 내에 존재하는 길이 n의 임의의 서브시퀀스 X = (x[0], x[1], ···, x[n-1])가 다음 조건을 만족하면, X를 Q와 유사하다고 간주하여 <S, S내 X의 오프셋>을 반환한다.

$$d(X,Q) \le \varepsilon, \text{ 여기서 } d(X,Q) = \sqrt{\sum_{i=0}^{n-1} (x[i]-q[i])^2}$$ 1)

2.2 FRM

참고 문헌[4]에서는 서브시퀀스 매칭을 위한 해결 방안을 제안하였다. 본 연구에서는 참고 문헌[8]에서 사용한 명칭을 따라 이 기법을 FRM이라 부른다. FRM에서는 미리 고정된 크기 w를 갖는 윈도우(window) 개념을 이용한다.

먼저, 각 시퀀스로부터 모든 가능한 위치에서 시작되는 길이 w의 슬라이딩 윈도우(sliding window)들을 추출하고, 각 윈도우를 이산 퓨리에 변환(discrete Fourier transform : DFT)을 이용하여 저차원 공간상의 점으로 변환한다. 본 연구에서는 이 점을 데이터 윈도우 점(data window point)이라 정의한다. 인덱싱의 대상이 되는 윈도우 점들의 수가 매우 많으므로, FRM에서는 이들을 다수의 점들을 포함하는 최소 포함 사각형(minimum bounding rectangle : MBR)들로 구성한 후, 이 MBR들을 다차원 인덱스(multidimensional index)의 하나인 R^*-트리[2]에 저장한다2).

서브시퀀스 매칭을 위해서는 질의 시퀀스로부터 크기 w의 디스조인트 윈도우(disjoint window)들을 추출하고, 윈도우들을 DFT하여 저차원 공간상의 윈도우 점들로 변환한다. 이를 질의 윈도우 점(query window point)이라 한다. 각 윈도우 점에 대하여 $\varepsilon/p^{1/2}$를 허용치로 갖는 범위 질의를 R^*-트리 상에서 수행한다. 여기서, p = ⌊ (질의 시퀀스 길이) / w ⌋ 이다. 이러한 범위 질의(range query)의 결과로 얻어진 데이터 윈도우 점들을 조사함으로써 최종 결과에 포함될 가능성이 높은 후보 서브시퀀스(candidate subsequence)들을 파악한다. 그 다음, 이 후보 서브시퀀스들을 디스크로부터 액세스하여 질의 시퀀스와의 유클리드 거리를 실제로 계산함으로써 최종적인 진위를 판단한다.

2.3 Dual-Match

FRM에서는 인덱싱을 위한 저장 공간의 오버헤드를 줄이

1) 즉, 유클리드 거리(Euclidean distance)가 주어진 허용치 ε 이하인 두 서브시퀀스 X와 Q를 유사하다고 간주한다.

2) 현재, 시계열 데이터베이스에서 가장 널리 사용되는 다차원 인덱스는 R^*-트리이다. 따라서 이후부터는 특별한 구분 없이 다차원 인덱스와 R^*-트리라는 용어를 혼용한다.

기 위하여 개별적 윈도우 점들 대신 다수의 윈도우 점들을 포함하는 MBR들을 R*-트리 내에 저장한다. 이러한 MBR 내부에는 죽은 공간(dead space)[2]이 존재하게 되므로, 이로 인하여 후보 서브시퀀스의 착오 채택이 발생되며, 이것은 처리 성능의 저하로 직결된다[8]. 참고문헌[8]에서는 이러한 문제점을 해결하기 위한 방법으로서 이원성 기반 서브시퀀스 매칭(duality-based subsequence matching : Dual-Match)을 제안하였다.

Dual-Match에서는 데이터 시퀀스로부터 슬라이딩 윈도우를 추출하고 질의 시퀀스로부터 디스조인트 윈도우를 추출하는 FRM과는 반대로 데이터 시퀀스로부터는 디스조인트 윈도우를 추출하고 질의 시퀀스로부터는 슬라이딩 윈도우를 추출하는 방식을 사용한다. 이와 같은 역할 교환을 통하여 Dual-Match에서는 R*-트리에 저장할 윈도우들의 수를 FRM의 약 1/w로 줄일 수 있다. 이 결과, MBR들을 저장하는 FRM과는 달리 Dual-Match에서는 윈도우 점 자체를 R*-트리에 저장하는 것이 가능해 진다. 따라서 MBR 내의 죽은 공간으로 인한 후보 서브시퀀스의 착오 채택이 발생되지 않으므로, 처리 성능이 크게 개선된다. 참고 문헌[8]의 성능 평가 결과에 의하면, Dual-Match의 검색 성능은 FRM과 비교하여 최대 430배까지 개선되는 것으로 나타났다.

3. 연구 동기

본 장에서는 FRM 및 Dual-Match에서 수행하는 공통적인 처리 과정을 요약하고, 후처리 과정(post-processing step)에서 발생하는 성능상의 문제점을 지적한다.

3.1 서브시퀀스 매칭 처리 과정

(그림 1)은 인덱스 검색과 후처리 과정으로 구성되는 서브시퀀스 매칭의 처리 과정을 나타낸 것이다. 큰 이등변 삼각형은 R*-트리를 나타내며, 아래쪽의 각 사각형은 시퀀스를 의미한다.

인덱스 검색은 각각의 질의 윈도우 점에 대하여 범위가 $\varepsilon/p^{1/2}$인 범위 질의를 R*-트리에 수행하는 과정이다. (그림 1)에서 점으로 채워진 네 개의 삼각형은 각 질의 윈도우 점을 이용하여 범위 질의를 처리할 때, R*-트리 내에서 액세스되는 부분을 나타낸다. 각 범위 질의의 결과는 그 범위 질의에서 사용된 질의 윈도우 점과 유클리드 거리가 $\varepsilon/p^{1/2}$ 이내인 데이터 윈도우들의 집합이다. 이들을 후보 윈도우(candidate window)라 정의한다. 이러한 처리를 하는 이론적인 배경은 "서로 다른 두 시퀀스의 유클리드 거리가 ε 이내이기 위한 필요 조건은 대응되는 윈도우 쌍 중 적어도 하나가 $\varepsilon/p^{1/2}$ 이내여야 한다"는 참고문헌[4]의 정리에 근거한다.

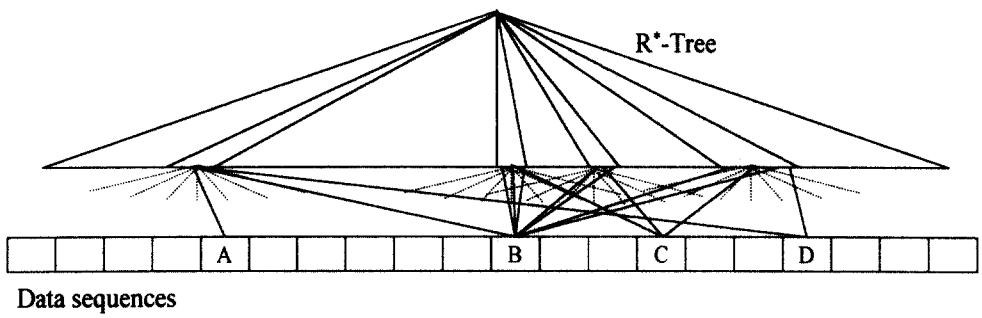
후처리 과정에서는 먼저 인덱스 검색에서 반환되는 각 후보 윈도우에 대하여 그 윈도우가 속하는 서브시퀀스를 파악한다. 이를 후보 서브시퀀스(candidate subsequence)라 정의한다. 그리고 나서, 이 후보 서브시퀀스가 포함되는 데이터 시퀀스를 디스크로부터 액세스하고, 후보 서브시퀀스와 질의 시퀀스와의 유클리드 거리를 실제로 계산함으로써 착오 채택(false alarm)[1]의 여부를 확인한다.

3.2 기존 기법들의 후처리 과정의 문제점

(그림 1)에서 나타난 바와 같이 R*-트리 내에서 윈도우 점들은 모두 독립적으로 관리된다. 또한, 기존의 FRM 및 Dual-Match의 후처리 과정에서는 ***인덱스 검색을 통하여 반환되는 순서***로 각 후보 윈도우가 포함되는 후보 서브시퀀스를 질의 시퀀스와 비교한다. 이러한 처리 방식은 다음과 같은 두 가지 측면에서 성능상의 문제들을 야기 시킨다.

3.2.1 디스크 액세스 오버헤드

이것은 동일한 데이터 시퀀스를 디스크로부터 반복적으로 액세스함으로써 발생하는 성능상의 문제이다. 이 문제는 인덱스 검색의 결과, 같은 데이터 시퀀스에 속하는 서로 다른 윈도우들이 후보 윈도우로 선택되는 경우에 발생한다. 즉, 같은 시퀀스에 속하는 후보 윈도우들이라 할 지라도 인덱스 검색의 결과로 반환되는 시점이 다르므로 같은 데이터 시퀀스를 디스크로부터 여러번 액세스해야 하는 것이다.

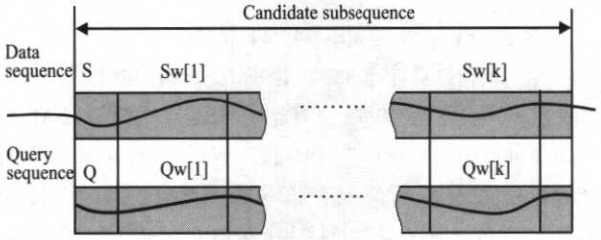


(그림 1) 서브시퀀스 매칭의 처리 과정

(그림 1)의 예에서 A, B, C, D는 모두 후보 윈도우를 포함하는 후보 시퀀스들이며, 이들은 후처리 과정에서 각각 1, 8, 3, 2회 디스크로부터 액세스되어야 함을 알 수 있다. 이와 같은 디스크로부터의 동일한 시퀀스의 중복 액세스는 전체 서브시퀀스 매칭의 성능을 저하시키는 중요한 원인이 된다.

3.2.2 CPU 처리 오버헤드

이것은 동일한 서브시퀀스를 질의 시퀀스와 두 번 이상 비교함으로써 발생하는 CPU 성능상의 문제이다. 이 문제는 인덱스 검색의 결과, 동일한 데이터 서브시퀀스에 속하는 서로 다른 윈도우들이 후보 윈도우로 선택되는 경우에 발생한다. (그림 2)의 경우를 예로 들어보자[3]. 질의 시퀀스 Q 내 Qw[1]에서 Qw[k]까지의 k개의 연속적인 윈도우들과 대응되는 데이터 시퀀스 S내 Sw[1]에서 Sw[k]까지의 윈도우들이 모두 후보 윈도우로 반환되면, 질의 시퀀스는 동일한 이 서브시퀀스와 k번 중복하여 비교하게 된다. 이러한 불필요한 중복 비교로 인하여 CPU 시간의 낭비를 초래한다.



(그림 2) 질의 시퀀스와 데이터 서브시퀀스의 비교.

4. 제안하는 기법

본 장에서는 제 3장에서 제시한 문제점들을 해결할 수 있는 새로운 후처리 기법을 제안한다. 제안하는 기법에서 추구하는 궁극적인 목표는 디스크 액세스 및 서브시퀀스 비교 과정에서 발생하는 중복을 완전히 제거하는 것이다. 제 4.1절에서는 본 논문에서 제안하는 해결 전략을 제시하고, 제 4.2절에서는 분석을 통하여 제안된 전략의 최적성을 규명한다.

4.1 해결 전략

중복의 문제가 발생하는 근본적인 원인은 인덱스 검색의 결과에 포함되는 후보 윈도우들이 무작위 순서대로 반환된다는 데 있다. 본 연구에서는 이러한 문제를 해결하기 위하여 반환되는 후보 윈도우를 위한 후처리 과정을 바로 수행하지 않고, 같은 시퀀스에 속하는 후보 윈도우들, 같은 서브시퀀스에 속하는 후보 윈도우들을 한꺼번에 처리하는 방식을 제안한다.

이를 위한 구체적인 방법은 다음과 같다. 먼저, 각 후보 윈도우에 대하여 대응되는 <seqID, subseqOffset>를 다중키(multi-attribute key)[5]로 사용하여 이진 트리(binary search tree)에 삽입한다. 여기서 seqID는 이 후보 윈도우가 속하는 데이터 시퀀스의 식별자이다. 또한, subseqOffset은 이 후보 윈도우에 의하여 후보로 추천된 seqID내 서브시퀀스의 시작 오프셋이다. 이 값은 이 후보 윈도우와 매치되는 질의 윈도우의 위치 정보를 이용하여 역으로 계산할 수 있다. 이러한 삽입 연산은 인덱스 검색 과정에서 후보 윈도우가 반환될 때마다 호출된다. 또한, 이러한 삽입 과정에서 삽입하고자 하는 키 값 <seqID, subseqOffset>이 이미 이진 트리 내에 존재하는 경우에는 이 키 값을 삽입하지 않고 무시한다. 예를 들어, (그림 2)에서 k개의 후보 윈도우들은 모두 같은 <S, offset>을 키 값으로 가지게 되므로 단 한번만 이진 트리에 삽입된다.

인덱스 검색 과정이 완료되면, 이진 트리 내에는 질의 시퀀스와 비교해야 할 후보 서브시퀀스들이 저장되어 있다. 후처리 과정에서는 이 이진 트리의 중위 순회(in-order traverse)를 통하여 얻어지는 후보 서브시퀀스 순서로 질의 시퀀스와의 비교를 수행한다.

4.2 성능 분석

후처리 과정에서 제안된 기법은 이진 트리의 중위 순회를 수행하므로, 이는 <seqID, subseqOffset>의 순서로 후보 시퀀스들을 액세스함을 의미한다. 따라서 같은 시퀀스에 속하는 후보 서브시퀀스들은 연속적으로 비교된다. 따라서 후보 서브시퀀스를 포함하는 시퀀스는 디스크로부터 단 한번만 액세스된다. 또한, 이진 트리 생성시 같은 시퀀스에 속하는 중복된 서브시퀀스들은 제거한바 있다. 따라서 동일한 서브시퀀스의 중복 비교는 발생하지 않는다. 제안된 기법은 반드시 필요한 디스크 액세스와 서브시퀀스 비교를 단 한번만 수행하므로 서브시퀀스 매칭의 후처리 과정을 위한 최적의 방법이다.

5. 성능 평가

본 장에서는 제안된 기법에 대한 정량적인 성능 평가 결과를 제시한다. 먼저, 제 5.1절에서는 성능 평가를 위한 실험 환경을 설명하고, 제 5.2절에서는 기존 기법과의 비교 실험을 통한 제안된 기법의 성능 개선 효과를 제시한다.

5.1 실험 환경

기존의 두 기법 중 더 나은 성능을 갖는다고 밝혀진 Dual-Match와의 성능 비교를 위하여 다음과 같은 실험을 수행하

3) 이 그림은 다른 현상을 설명하기 위하여 참고문헌[8]에서 사용하였던 것이다. 본 논문에서는 CPU 처리의 오버헤드를 설명하기 위하여 이 그림을 그대로 인용하였다.

였다. 먼저, 실험에서 사용된 데이터는 1994년 11월부터 1998년 5월까지 수집한 한국의 주식 데이터 KStock이다. KStock은 길이 1024의 620개 종목의 정규화된 데이터 시퀀스들로 구성된다. 인덱싱을 위하여 사용된 윈도우의 길이는 30, 60, 90의 세 가지이며, 사용된 질의 시퀀스로는 KStock내의 하나의 데이터 시퀀스를 무작위로 선정하여 임의의 위치에서 길이 200만큼을 추출하였다. 또한, 질의를 위한 유사 허용치 ε으로는 21개의 최종 질의 결과를 반환하도록 하는 2.0를 사용하였다. 인덱싱을 위하여 추출된 특징은 각 윈도우를 DFT하여 얻어진 계수 중 큰 에너지를 갖는 앞쪽의 4개이며, 이 결과 4차원 R^*-트리를 구성하였다.

5.2 실험 결과

<표 1>은 제안하는 기법과 기존 기법에서 각각 발생하는 시퀀스 액세스 횟수를 (1) 최종 결과로 반환되는 것(true answers), (2) 착오 채택되는 것(false alarms), 그리고 (3) 두 가지 모두를 포함한 것(total) 등 세 가지 측면에서 비교한 것이다. 전체 결과를 나타내는 마지막 두 행에서와 같이, 제안된 기법이 윈도우의 크기에 따라 55배에서 156배까지의 성능 개선 효과를 가지는 것으로 나타났다. 또한, 윈도우가 작을 수록 이러한 성능 개선 효과가 더욱 커지는 것으로 나타났다. 이것은 윈도우가 작아질수록 착오 채택의 수가 많아짐에 따라 후보 시퀀스를 액세스하는 회수가 많아지기 때문이다.

<표 2>는 제안하는 기법과 기존 기법에서 각각 발생하는 서브시퀀스 비교 횟수를 (1) 최종 결과로 반환되는 것(true answers), (2) 착오 채택되는 것(false alarms), 그리고 (3) 두 가지 모두를 포함한 것(total)등 세 가지 측면에서 비교한 것이다. 전체 결과를 나타내는 마지막 두 행에서와 같이, 제안된 기법이 윈도우의 크기에 따라 37%까지의 성능 개선 효과를 가지는 것으로 나타났다. 또한, 이러한 성능 개선 효과는 윈도우가 작을 수록 커지는 것으로 나타났다. 이것은 윈도우가 작아질수록 동일한 서브시퀀스 내에 속하는 후보 윈도우 수가 많아짐에 따라 Dual-Match에서는 동일한 후보 서브시퀀스를 질의 시퀀스와 비교하는 회수가 많아지기 때문이다.

〈표 1〉 디스크 액세스 회수의 비교

	방법 \ w	30	60	90
True Answers	ours	21	21	21
	Dual-Match	108	50	21
False Alarms	ours	532	373	250
	Dual-Match	86012	27241	14875
total	ours	553	394	271
	Dual-Match	86210	27291	14896

〈표 2〉 서브시퀀스 비교 회수의 비교

	방법 \ w	30	60	90
True Answers	ours	21	21	21
	Dual-Match	108	50	21
False Alarms	ours	62602	24434	14548
	Dual-Match	86012	27241	14875
total	ours	62623	24455	14569
	Dual-Match	86210	27291	14896

실험 결과, 디스크 액세스 회수 측면에서의 제안된 기법의 성능 개선 효과가 서브시퀀스 비교 측면과 비교하여 훨씬 큰 것으로 나타났다. 디스크 액세스 비용이 전체 처리 시간의 대부분을 차지한다는 것을 고려하면, 이러한 결과는 매우 바람직한 것이다.

6. 결 론

본 연구에서는 서브시퀀스 매칭의 후처리 과정에서 발생하는 기존 기법들의 문제점을 지적하고, 이를 해결할 수 있는 최적의 기법을 제안하였다. 제안된 기법은 이진 트리 내에 후보 시퀀스에 대한 정보를 삽입해 둠으로써 같은 시퀀스에 속하는 후보 윈도우들과 같은 서브시퀀스에 속하는 후보 윈도우들을 함께 처리하는 방식을 사용한다. 이 결과, 반드시 필요한 디스크 액세스와 서브시퀀스 비교를 단 한 번씩만 수행한다. 따라서 제안된 기법은 서브시퀀스 매칭의 후처리 과정을 위한 최적의 방법이다. 제안된 기법의 성능 개선 효과를 검증하기 위하여 실제 주식 데이터를 위한 성능 평가를 수행하였다. 실험 결과에 의하면, 제안된 기법은 기존의 기법과 비교하여 55배에서 156배까지의 성능 개선 효과가 있는 것으로 나타났다.

참 고 문 헌

[1] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient Similarity Search in Sequence Databases," In Proc. Int'l. Conf. on Foundations of Data Organization and Algorithms, FODO, pp.69-84, Oct., 1993.

[2] N. Beckmann et al., "The R^*-tree : An Efficient and Robust Access Method for Points and Rectangles," In Proc. Int'l. Conf. on Management of Data, ACM SIGMOD, pp.322-331, May, 1990.

[3] M. S. Chen, J. Han, and P. S. Yu, "Data Mining : An Overview from Database Perspective," IEEE Trans. on Knowledge and Data Engineering, Vol.8, No.6, pp.866-883, 1996.

[4] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-series Databases," In Proc. Int'l. Conf. on Management of Data, ACM SIGMOD, pp. 419-429, May, 1994.

[5] J. Gray and A. Reuter, Transaction Processing : Concepts and Techniques, Morgan Kaufman Publishers, 1993.
[6] S. W. Kim, S. H. Park, and W. W. Chu, "An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," In Proc. IEEE Int'l. Conf. on Data Engineering, IEEE ICDE, pp.607-614, 2001.
[7] 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 Proc. ACM Int'l. Conf. on Information and Knowledge Management, ACM CIKM, pp.480-487, 2000.
[8] Y. S. Moon, K. Y. Whang, and W. K. Loh, "Duality-Based Subsequence Matching in Time-Series Databases," In Proc. IEEE Int'l Conf. on Data Engineering, IEEE ICDE, pp. 263-272, 2001.
[9] S. H. Park, S. W. Kim, and W. W. Chu, "Segment-Based Approach for Subsequence Searches in Sequence Databases," In Proc. ACM Int'l. Symp. on Applied Computing, ACM SAC, pp.248-252, 2001.
[10] D. Rafiei and A. Mendelzon, "Similarity-Based Queries for Time-Series Data," In Proc. Int'l. Conf. on Management of Data, ACM SIGMOD, pp.13-24, 1997.

김 상 욱

e-mail : wook@kangwon.ac.kr
1989년 서울대학교 컴퓨터공학과 졸업(학사)
1991년 한국과학기술원 전산학과 졸업(석사)
1994년 한국과학기술원 전산학과 졸업(박사)
1991년 미국 Stanford University, Computer Science Department, Summer Intern
1994년~1995년 정보전자연구소 Post-Doc.
1999년~2000년 미국 IBM T.J. Watson Research Center, Post-Doc.
1995년~현재 강원대학교 컴퓨터정보통신공학부 부교수
관심분야 : 시퀀스 매칭, DBMS, 저장 시스템, 주기억장치 DBMS, 임베디드 DBMS