주식 데이타베이스에서 질의간 따름 관계를 이용한 연속 질의의 처리

(Continuous Query Processing Utilizing Follows Relationship between Queries in Stock Databases)

하 유 민 † 김 상 욱 †† 박 상 현 †††
(You-min Ha) (Sang-Wook Kim) (Sanghyun Park)

요 약 본 논문에서는 주식 데이타베이스로부터 탐사된 다수의 규칙들을 이용하여 주식 투자 추천을 요구하는 대량의 연속 질의들을 효과적으로 처리하는 방안에 관하여 논의한다. 먼저, 본 논문에서는 주식 투자 추천을 위한 사용자 질의의 특성을 분석함으로써 질의간에 존재하는 새로운 관계인 '따름 관계'를 정의한다. 두 질의 Q_1, Q_2간의 추천값 X에 대한 따름 관계는 '만일 선행 질의 Q_1의 추천값이 X이면, 추종 질의 Q_2의 추천값은 항상 X인 관계'를 의미한다. 이러한 따름 관계가 존재하는 경우, 추종 질의 Q_2의 추천값은 선행 질의 Q_1의 추천값을 이용하여 바로 결정할 수 있으므로 Q_2를 위한 질의 처리 과정을 제거할 수 있다. 본 논문에서는 전체 사용자 질의들간의 따름 관계들을 파악하여 그래프 형태로 표현하는 방법을 제안한다. 또한, 처리 과정이 제거되는 질의들의 수가 최대가 되도록 이러한 그래프를 탐색하여 질의 처리 순서를 결정하는 방법을 제안한다. 따름 관계를 기반으로 하는 제안된 방식을 이용하는 경우, 많은 사용자 질의들은 실제 질의 처리 과정이 불필요하게 되므로 전체 시스템의 처리 성능을 크게 개선할 수 있다. 실제 주가 데이타를 이용한 실험을 통하여 제안한 질의 처리 방식의 우수성을 규명한다. 실험 결과에 의하면, 제안된 방식에 의한 전체 질의 처리 시간은 기존 방식에 의한 시간의 10%이하로 줄어드는 것으로 나타났다.

키워드 : 연속 질의 처리, 주식 데이타베이스, 규칙 탐사

Abstract This paper analyzes the properties of user query for stock investment recommendation, and defines the 'following relation', which is a new relation between two queries. A following relation between two queries Q_1, Q_2 and a recommendation value X means 'If the recommendation value of a preceding query Q_1 is X, then a following query Q_2 always has X as its recommendation value'. If there exists a following relation between Q_1 and Q_2, the recommendation value of Q_2 is decided immediately by that of Q_1, therefore we can eliminate the running process for Q_2. We suggest two methods in this paper. The former method analyzes all the following relations among user queries and represents them as a graph. The latter searches the graph and decides the order of queries to be processed, in order to make the number of eliminated query-running process maximized. When we apply the suggested procedures that use the following relation, most of user queries do not need to be processed directly, hence the performance of running overall queries is greatly improved. We examined the superiority of the suggested methods through experiments using real stock market data. According to the results of our experiments, overall query processing time has reduced less than 10% with our proposed methods, compared to the traditional procedure.

Key words : Continuous Query Processing, Stock Database, Rule Discovery

· 본 논문은 제주대학교를 통한 정보통신부 및 정보통신진흥원의 대학IT연구센터 지원사업(IITA-2005-C1090-0502-0009) 및 서울시가 시행하고 서울시립대학교 "지능형 도시 사업단(스마트-유비쿼터스-시티 사업단)"이 주관하는 "스마트시티를 위한 지능형 도시정보 컨버전스 시스템 개발"사업(10561)에서 지원을 받았습니다.

† 학생회원 : 연세대학교 컴퓨터과학과
ymha@cs.yonsei.ac.kr
†† 종신회원 : 한양대학교 정보통신학부 교수
wook@hanyang.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
sanghyun@cs.yonsei.ac.kr

논문접수 : 2006년 8월 7일
심사완료 : 2006년 10월 30일

1. 서 론

1.1 주식 추천의 필요성

시계열 데이타(time-series data)란 시간의 흐름에 따라 일정 간격으로 객체의 변화를 관측하여 얻어진 값들의 리스트이다[1-4]. 이러한 시계열 데이타는 시간에 따르는 경제 현상이나 자연 현상의 변화를 나타내며, 임의의 한 시점에서 관측된 값은 그 이전까지의 누적된 값들로부터 영향을 받게 된다[5]. 따라서 시계열 데이타를 분석함으로써 과거에 관측된 값들로부터 규칙성을 발견하고, 이를 모델링하여 미래에 관측될 값을 예측할 수 있다.

주가의 변화를 기록한 데이타는 대표적인 시계열 데이타의 하나이다[6-8]. 주식 투자자의 궁극적인 목적은 수익률을 극대화하는 것이다. 주가 데이타의 분석을 통하여 지수 흐름, 주가의 변화 시점, 거래 시세 등을 예측하여 주식의 매매 시점을 잘 선택할 수 있다면 성공적인 주식 투자를 기대할 수 있을 것이다. 그러나 주가 변동이 빈번하게 일어나는 상황에서 이익을 낼 수 있는 주식의 매수나 매도 시점을 투자자가 직접 결정하는 것은 그리 쉬운 일이 아니다[9].

각 주식 투자자가 원하는 투자 조건은 매우 다양하다. 어떤 투자자는 손실 위험이 크더라도 많은 수익을 얻을 수 있는 공격적인 투자를 원할 수 있으며, 반면 어떤 투자자는 적은 수익을 얻더라도 손실을 최소화하는 투자를 원할 수도 있다. 따라서, 투자자가 원하는 매수/매도 조건을 직접 설정하면, 시스템이 이 조건이 만족되는 시점을 포착하여 해당 주식을 투자자에게 자동으로 추천해 주는 주식 투자 모델이 필요하다.

1.2 이전 연구 요약

참고문헌 [10]에서는 전술한 요건들을 만족시키는 주식 투자 시스템을 제안하였다. 이 연구에서 제안한 규칙 모델에서, 규칙은 규칙 헤드(rule head)와 규칙 바디(rule body)로 구성되어 있다. 주가 데이타에서 빈번하게 발생하는 패턴을 발견하고, 각 패턴을 지지하는 과거의 주가 데이타를 참조하여 해당 빈번 패턴 발생 이후의 변화 경향을 예측한다. 이때, 각각의 사용자가 자신이 원하는 투자 조건을 반영한 질의를 입력할 수 있다.

1.3 연구 동기

참고문헌 [10]에서는 이처럼 규칙 모델을 정의하고, 규칙 매칭 질의를 효과적으로 처리하는 것을 목표로 삼았다. 그러나 이 규칙 모델이 적용되는 주식 투자 환경에서는 매우 많은 투자자가 존재하며, 각 투자자가 여러 관심 종목에 대한 여러 개의 질의를 입력하게 된다. 또한, 거래가 발생함에 따라 모든 종목의 주가는 대단히 빈번하게 갱신된다. 이러한 주가 데이타 갱신이 발생할 때마다 투자자들이 입력한 규칙 매칭 질의들을 모두 다시 실행함으로써 각 질의의 추천값을 다시 계산해야 한다. 이와 같이, 데이타의 변경에 의하여 동일한 질의를 지속적으로 처리해야 하는 것을 연속 질의 처리(continuous query processing) 라 한다[11]. 연속 질의 처리 환경에서는 전체 질의를 모두 처리하는 데에 드는 비용을 비약적으로 감소시키는 방안이 필요하다.

본 논문에서는 주식 데이타베이스로부터 탐사된 다수의 규칙들을 이용하여 주식 투자 추천을 요구하는 대량의 연속 질의들을 효과적으로 처리하는 방안에 관하여 논의한다. 먼저, 본 논문에서는 주식 투자 추천을 위한 사용자 질의의 특성을 분석함으로써 질의간에 존재하는 새로운 관계인 '따름 관계'를 정의한다. 두 질의 Q_1, Q_2 간의 추천값 X에 대한 따름 관계는 '만일 선행 질의 Q_1의 추천값이 X이면, 추종 질의 Q_2의 추천값이 항상 X인 관계'를 의미한다. 이러한 따름 관계가 존재하는 경우, 선행 질의 Q_1의 추천값을 이용하여 추종 질의 Q_2의 추천값을 바로 결정할 수 있으므로, Q_2를 위한 질의 처리 과정을 생략할 수 있다. 본 논문에서는 질의들간의 따름 관계를 파악하여, 전체 질의들을 그래프 형태로 표현하는 방법을 제안한다. 또한, 이러한 그래프를 탐색함으로써 처리 과정이 제거되는 질의들의 수가 최대가 되도록 질의 처리 순서를 결정하는 방법을 제안한다. 따름 관계를 기반으로 하는 이 방법을 이용하는 경우, 많은 질의들의 처리 과정이 불필요하게 되므로 전체 시스템의 처리 성능을 크게 개선할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 이전 연구에서 제안한 규칙 모델에 관하여 간략히 설명한다. 3장에서는 본 연구에서 제안하는 시스템이 실행되는 환경의 여러 가지 제약 조건으로 인해 발생하는 문제들을 정의한다. 4장에서는 이러한 문제들을 해결하는 방법을 제안한다. 5장에서는 제안한 방법의 우수성을 실험으로 검증한다. 끝으로, 6장에서 본 논문을 요약하고 결론을 내린다.

2. 주식 투자 추천 시스템

이 장에서는 참고문헌 [10]에서 제안된 규칙 모델과 질의 모델에 대하여 간략하게 설명한다.

2.1 규칙 모델

규칙 모델에서 규칙은 규칙 헤드(rule head)와 규칙 바디(rule body)로 구성된다. 규칙 헤드는 시간에 따라 변하는 주가 데이타에서 빈번하게 발견된 패턴이며, 규칙 바디는 일정 시간 간격이 지난 후 주가 변화 양상을 가리킨다. 이를 좀더 명확히 정의하면 다음과 같다.

$$H \xrightarrow{t} B(s,c)$$

여기서, H는 규칙 헤드이며, B는 규칙 바디이다. 이 규칙은 H에 해당되는 사건이 발생한 후, t 시간이 흐른 후에는 B에 해당되는 사건이 발생하였음을 의미한다.

주가 변화의 패턴이 규칙으로서 가치를 가지기 위해서

는 과거에 발생하였던 많은 패턴들이 규칙과 부합하여야 한다. s는 아래와 같이 정의되는 지지도(support)로서 H에 해당되는 패턴 P가 과거에 발생하였던 상대 빈도를 표현한다. 즉, 규칙 헤드 H와 매치하는 실제 주가 변화 패턴이 얼마나 많이 발생하였는가를 나타내는 척도이다.

$$s(H)=\frac{H\text{와 매치되는 패턴들의 발생 수}}{H\text{와 매치되는 패턴과 길이가 동일한 모든 패턴의 발생 수}}\times 100$$

또한, 규칙으로서 가치를 가지기 위해서는 H와 매치하는 과거 패턴들이 B 구간 내에서 일정한 경향을 보여야 한다. c는 아래와 같이 정의되는 신뢰도(confidence)로서 H와 매치하는 과거 패턴들 중 얼마나 많은 수가 규칙 바디 B를 위한 조건을 만족시키는가를 표현한다.

$$c(H,B)=\frac{H\text{와 매치되며, }B\text{의 조건을 만족시키는 패턴의 발생 수}}{H\text{와 매치되는 패턴의 발생 수}}\times 100$$

본 논문에서 제안하는 기법에서는 과거의 주가 데이타베이스를 분석함으로써 지지도와 신뢰도가 사전에 지정된 값 이상인 규칙들을 탐사하고, 투자자의 관심 종목의 최근 주가 변화 패턴이 탐사된 어떤 규칙의 헤드 H와 매치됨이 발견되면, 해당 규칙의 바디 B를 참조하여 해당 종목에 대한 투자 유형을 투자자에게 추천한다. 투자 유형은 '매수', '매도', '보유', '무추천' 등이 있을 수 있다. 투자 유형은 규칙 바디에 의하여 결정되며, 규칙 바디에 대한 조건은 투자자의 투자 성향에 따라 달라질 수 있다.

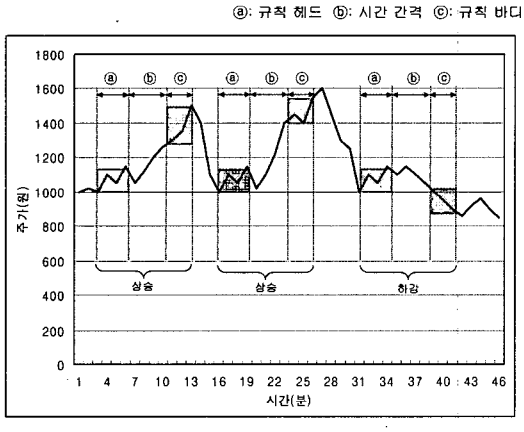


그림 1 규칙 모델 예제

예를 들어, 그림 1은 어떤 종목의 주가 변화를 나타낸 것이다. 이 종목의 주가 데이타에서 ⓐ 형태의 패턴이 3회 발생하였고, 이 패턴이 발생한 이후 일정한 시간 ⓑ 만큼 지난 후의 규칙 바디 ⓒ에서 주가가 오른 횟수가 2회, 내린 횟수가 1회였다고 하자. 이로 미루어, 다음에 이와 같은 패턴이 다시 발생한다면 주가가 오를 확률이 높으므로, 예측값 'BUY'를 추천하게 된다. 위의 그림에서는 3회 발생한 패턴을 예로 들어 설명하였으나, 실제로는 현재까지의 주가 변화 데이타 중 최소 지지도 이상의 발생 빈도를 갖는 빈번 발생 패턴(frequent pattern)[12,13] 들을 규칙 헤드로 사용한다.

본 응용에서 H는 그림 1에 나타난 예제의 ⓐ 구간에서와 같은 특정 주가 변화 패턴 P의 발생과 대응되는 사건이다. 또한, B는 ⓒ 구간 내에서 발생하는 주가의 특성을 요약하는 사건이다. 예를 들어, 위의 그림 1의 예제에서 B는 "상승"으로 표현될 수 있다. 투자자는 자신이 추천 받기를 원하는 투자 유형과 관련하여 이 ⓒ 구간 내 주가 특성에 관한 구체적인 조건을 명시할 수 있다. 이를 규칙 바디의 조건이라 명명한다. 이 조건은 구간 내 주가 특성이 어떠한 경향을 보일 때 이를 상승으로 간주할 것인가 하는 조건을 나타낸다. 위의 예제에서 투자자는 ⓐ 구간의 마지막 주가 대비 ⓒ 구간에서의 평균 주가 상승률이 10% 이상 되는 것을 규칙 바디의 조건으로 설정할 수 있으며, 이 경우 그림 1의 주가 변화 형태는 의미 있는 규칙으로 생성될 수 있다. 이와 같이, 이러한 규칙 바디의 조건은 투자자의 성향에 따라 달라질 수 있다.

주가 데이타는 실수이므로 빈번 발생 패턴이 발생할 가능성은 매우 낮다. 따라서 주가 변화율의 도메인을 다수의 구간들로 나누어, 실수값인 각 주가 변화율을 구간과 대응되는 문자로 변환한 후, 이로부터 빈번 발생 패턴을 탐색하는 방법을 사용한다. 탐색된 빈번 발생 패턴들은 매번 주가가 갱신될 때마다 다수의 질의들에 대하여 빠르게 검색되어야 하므로, 이들에 대한 인덱스를 구성하여 저장한다.

각 투자자는 자신의 투자 성향에 따라 주식 종목, 그 종목을 매도할 시점과 매수할 시점을 결정하는 주가 변화율의 최소/최대값, 예측할 구간의 길이 등을 정하여 질의를 작성한다. 이 값들은 해당 질의에 대한 규칙 바디의 특성을 결정한다.

2.2 질의 모델

[정의 1] 질의 Q
투자 추천을 요구하기 위하여 투자자가 정의하는 질의 Q의 형태는 다음과 같다.

$$Q = (I, T, BL, [\alpha, \beta], mC)$$

각각의 변수는 다음과 같은 의미를 가진다.
- I : 예측하려는 종목.
- T : 규칙 헤드와 규칙 바디 사이의 시간 간격.
- BL : 규칙 바디의 길이.
- [α, β] : 보유 변동률이며, α와 β의 의미는 정의 2에서 설명한다.
- mC : 최소 신뢰도이며, mC > 0.5이다. 그 의미는 정의 2에서 설명한다. □

종목 I의 주가가 변화할 때마다 Q를 수행하며, 현재

까지의 주가 변화가 빈번 발생 패턴과 매치되었을 때 해당 종목에 대한 추천값을 반환한다. 이와 같이, 데이타의 변경에 의하여 동일한 질의를 지속적으로 처리해야 하는 것을 연속 질의 처리(continuous query processing)라 한다.

질의 Q의 실행 결과 F(Q)는 다음과 같은 값을 가진다.

[정의 2] 질의 Q의 실행 결과 F(Q)
빈번 발생 패턴이 발생한 각 사례(case)에 대하여, 규칙 헤드의 마지막 주가에 대한 규칙 바디의 주가 평균값의 증가 비율을 r이라 하면, 질의 Q=(I, T, BL, [α, β], mC)의 실행 결과 F(Q)는 다음과 같이 정의된다.

F(Q) = X, X ∈ { SELL, HOLD, BUY, NONE }

이며, α는 'HOLD' 선택 하한선, β는 'HOLD' 선택 상한선이다. 이때, X의 결과로 올 수 있는 4개의 값들을 추천값이라 하며, 각각 다음과 같은 경우에 질의 Q의 추천값으로 결정된다.

- SELL: 종목 I의 모든 빈번 발생 패턴에 대하여, $r \le \alpha$ 인 경우의 비율이 mC 이상일 때.
- HOLD: 종목 I의 모든 빈번 발생 패턴에 대하여, $\alpha < r < \beta$인 경우의 비율이 mC 이상일 때.
- BUY: 종목 I의 모든 빈번 발생 패턴에 대하여, $r \ge \beta$ 인 경우의 비율이 mC 이상일 때.
- NONE: 종목 I의 모든 빈번 발생 패턴에 대하여, SELL, HOLD, BUY 어느 결과의 비율도 mC를 넘지 못했을 때.

이때, SELL, HOLD, BUY 3가지의 추천값 중 2개 이상의 추천값이 동시에 결정되는 경우를 방지하기 위하여, 정의 1에서와 같이 최소 신뢰도 mC는 0.5보다 큰 값을 입력하도록 제한을 둔다. 따라서, 한 시점에서 F(Q)는 유일한 값을 추천한다. F(Q)의 값이 X인 경우, F(Q)=X 로 표시하고, '질의 Q를 실행한 결과, 추천값은 X이다' 라고 읽는다. □

질의 Q를 결정하는 모든 변수값들은 투자자가 원하는 대로 등록할 수 있다. 따라서, 이 질의 모델은 투자자들의 다양한 성향을 유연하게 수용할 수 있다는 장점을 가진다. 실험 결과 이 주식 투자 추천 시스템은 70% 이상의 예측 정확도를 가지는 것으로 나타났다[10].

3. 문제 정의

이 장에서는 본 연구에서 대상으로 하는 시스템의 환경을 고찰하고, 환경의 특성으로 인하여 발생하는 문제들을 정의한 다음, 간단한 예를 통하여 문제를 해결하는 핵심 아이디어를 제시한다.

먼저, 규칙 헤드의 탐사 대상이 되는 원본 주가 데이타로는 장기간 수집된 데이타를 사용한다. 이 원본 주가 데이타에서 규칙 헤드로 사용되는 빈번 발생 패턴들을 탐사하여 저장하며, 효과적인 질의 처리를 위하여 이 규칙 헤드들에 대한 인덱스를 구성한다. 본 연구에서는 원본 주가 데이타에 새롭게 추가된 주가를 인덱스에 반영하는 것은 고려하지 않기로 한다.

한편, 빈번 발생 패턴과의 매칭을 위한 각 종목의 현재 주가는 주기적으로 갱신된다. 그리고 질의에 포함된 각 종목의 주가가 갱신될 때마다 이 질의가 수행된다. 이 때, 빈번 발생 패턴이 저장된 규칙 헤드 데이타베이스에서 해당 종목의 현재까지의 주가 데이타와 매치되는 규칙 헤드가 존재하는지를 검색하고, 만약 존재할 경우 그 종목에 대한 투자 추천 여부를 결정한다. 규칙 바디를 결정하는 조건이 질의마다 다르기 때문에, 규칙 헤드가 발생한 수만큼 디스크에서 주가를 읽어 변화율을 계산해야 한다. 이로 인해 질의 하나를 수행할 때마다 디스크에 많은 수의 임의 접근이 발생하게 된다.

또한, 이 시스템이 적용되는 주식 투자 환경에서는 매우 많은 수의 투자자들이 존재한다. 또한, 한 투자자가 관심을 가지는 종목의 수가 여러 개이므로, 다수의 질의를 입력할 수 있다. 투자자들이 입력한 모든 질의는 주메모리에 저장하여 실행하며, 질의 데이타를 잃지 않기 위하여 디스크에 질의들을 백업한다. 수많은 투자자가 다수의 질의들을 입력하는 상황이므로, 실행해야 하는 전체 질의의 개수는 대단히 많다. 또한, 이렇게 많은 질의들이 주가가 갱신될 때마다 주기적으로 수행된다. 이러한 문제는 스트림 데이타 처리 연구에서 중요하게 다루어지는 주제인 연속 질의 처리 문제의 일종이다.

Q_1=('삼성전자', 0, 1, [-0.01, 0.01], 0.7), Q_2=('삼성전자', 0, 1, [-0.02, 0.02], 0.6)인 두 질의 Q_1과 Q_2에서, 종목과 시간 간격, 규칙 바디의 길이는 모두 동일한 것을 알 수 있다. 이때, $F(Q_1)$=HOLD 였다면, 이것은 종목 '삼성전자'의 모든 빈번 발생 패턴에 대하여 질의 Q_1을 실행한 결과, -0.01< 주가 증가율 <0.01 인 경우의 비율이 70% 이상이었다는 것을 의미한다. 그런데 Q_1과 Q_2의 종목, 시간 간격, 규칙 바디의 길이가 같으므로, Q_1을 실행했을 때와 동일한 시점에서 질의 Q_2를 실행할 때에 검토해야 하는 규칙 헤드와 규칙 바디는 질의 Q_1을 실행할 때와 완전히 동일하다. 질의 Q_1을 실행한 결과, 주가 증가율이 -0.01과 0.01 사이에 있었던 경우가 70% 이상이었으므로, 주가 증가율이 -0.02와 0.02 사이에 있었던 경우가 60% 이상이었을 때 HOLD를 추천하게 되는 Q_2의 실행결과 $F(Q_2)$는 Q_2를 실제로 수행하지 않고도 명백히 HOLD임을 알 수 있다.

다수의 질의들을 주가 데이타 입력 주기마다 반복해서 모두 처리해야 하므로, 전체 질의들을 빠르게 처리하는 방안이 필요하다. 본 연구에서는 이러한 요구를 만족시키기 위하여, 전술한 바와 같은 서로 다른 질의 Q_1과 Q_2간의 관계에 대하여 주목한다.

4. 따름 관계를 이용한 질의 처리

이 장에서는 질의 사이에 존재하는 따름 관계에 대하여 논의한다. 정의와 증명을 통하여 따름 관계를 고찰하며, 이를 이용하여 직접 처리되는 질의의 수를 줄이는 방법에 대하여 설명한다.

4.1 따름 관계의 정의

이 절에서는 질의들간의 따름 관계를 정의하고, 이러한 따름 관계 발생을 위한 조건을 제시한다.

[정의 3] 따름 관계와 따름 관계 집합
질의 Q_1과 Q_2, 추천값 X에 대하여, 따름 관계 R은 다음과 같이 정의된다.

$$R(Q_1, Q_2, X) \equiv \text{'만약 } F(Q_1)=X \text{ 이면, } F(Q_2)=X \text{ 이다.'}$$

이때, $R(Q_1, Q_2, X)$는 'Q_2는 추천값 X에 대하여 Q_1을 따른다' 고 읽으며, Q_1을 선행 질의(preceding query), Q_2는 추종 질의(following query)라고 정의한다. 또한, 추천값 X에 대하여 질의 Q_1을 선행 질의로 하는 모든 따름 관계들을 모은 집합을 따름 관계 집합이라 하며, $RS(Q_1, X)$라 표기한다. □

질의 Q=(I, T, BL, [α, β], mC) 가 있을 때, 종목 I의 새로운 주가가 발생할 때마다 질의 Q를 수행하고, 그 결과를 반환하게 된다. 종목 I에 대한 여러 개의 질의가 존재할 때, 같은 종목 I에 대한 질의들 사이에는 다음과 같은 따름 관계가 존재한다.

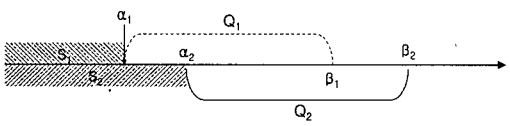
[정리 1] 따름 관계 $R(Q_1, Q_2, X)$
종목 I, 시간 간격 T, 규칙 바디의 길이가 모두 같은 질의 Q_1과 Q_2가 존재할 때, Q_1과 Q_2는 다음과 같이 표현할 수 있다.

$$Q_1 = (I, T, BL, [\alpha_1, \beta_1], mC_1)$$
$$Q_2 = (I, T, BL, [\alpha_2, \beta_2], mC_2)$$

이때, 다음과 같은 세 가지 따름 관계가 존재한다.

1.1. $\alpha_1 \le \alpha_2$ 이고, $mC_1 \ge mC_2$ 이고, $F(Q_1)$ = SELL 이면, 항상 $R(Q_1, Q_2, SELL)$ 이다.
1.2. $\alpha_1 \ge \alpha_2$ 이고, $\beta_1 \le \beta_2$ 이고, $mC_1 \ge mC_2$ 이고, $F(Q_1)$ = HOLD 이면, 항상 $R(Q_1, Q_2, HOLD)$ 이다.
1.3. $\beta_1 \ge \beta_2$ 이고, $mC_1 \ge mC_2$ 이고, $F(Q_1)$ = BUY 이면, 항상 $R(Q_1, Q_2, BUY)$ 이다.

[증명]
먼저, 정리 1.1을 증명해 보자. Q_1과 Q_2의 보유 변동률 $[\alpha_1, \beta_1]$과 $[\alpha_2, \beta_2]$를 그림으로 나타내면 다음과 같다. 이때, 가정에 의해 $\alpha_1 \le \alpha_2$ 이다.



빈번 발생 패턴의 마지막 주가와 그로부터 T시간 뒤의 BL 길이 구간 내의 평균 주가를 비교한 주가 변화율이 α_1 보다 작으면 $F(Q_1)$=SELL 인 사례(case)가 된다. 빈번 발생 패턴의 발생 횟수를 n이라 할 때, $F(Q_1)$=SELL이 추천된 사례의 수 $r_1(SELL)$에 대한 신뢰도 $C_1(SELL)$은 다음과 같이 구할 수 있다.

$$C_1(SELL) = r_1(SELL) \div n$$

이때, $C_1(SELL) \ge mC_1$ 이면 $F(Q_1)$ = SELL 이 된다.

Q_1을 실행한 결과 $F(Q_1)$이 SELL 이라고 하자. 정의에 의해 영역 S_1에 속하는 사례 $r_1(SELL)$ 의 수는,

$$r_1(SELL) = n \times C_1(SELL) \ge n \times mC_1$$

이다. 그림에 나타난 바와 같이 $S_1 \subseteq S_2$ 이므로, S_2에 속하는 사례 $r_2(SELL)$ 의 수는

$$r_2(SELL) \ge r_1(SELL)$$

이다. 따라서,

$$r_2(SELL) \ge r_1(SELL) \ge n \times mC_1 \ge n \times mC_2$$

가 된다.
C_1의 정의와 마찬가지로, $C_2(SELL) = r_2(SELL) \div n$ 이므로,

$$C_2(SELL) = r_2(SELL) \div n \ge n \times mC_2 \div n = mC_2$$

이다. $C_2(SELL) \ge mC_2$ 이므로, 정의에 의하여 $F(Q_2)$=SELL 이 된다.
따라서, $\alpha_1 \le \alpha_2$ 이고, $mC_1 \ge mC_2$ 이며, $F(Q_1)$=SELL 이면 $R(Q_1, Q_2, SELL)$이 항상 성립한다.

정리 1.2와 1.3도 유사한 방법으로 증명할 수 있다. □

정리 1에서 증명한 바와 같이 따름 관계를 가지는 다수의 질의들이 존재하며, 선행 질의의 추천값에 의하여 추종 질의의 추천값을 자동으로 결정할 수 있으므로, 고비용의 질의 처리 과정을 생략할 수 있다.

4.2 질의 집합의 구성 방법

질의간의 따름 관계는 유향 그래프(directed graph)로 표현할 수 있다. 그래프의 각 정점은 질의를 나타내며, 화살표는 추천값에 대한 따름 관계를 나타낸다. 예를 들어, 그림 2는 $R(Q_1, Q_2, BUY)$를 유향 그래프로 표현한 것이다.

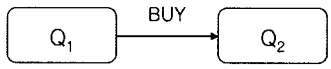


그림 2 따름 관계 $R(Q_1, Q_2, BUY)$

n개의 질의가 있을 때, 한 질의로부터 나갈 수 있는 화살표는 조건에 따라 최대 n-1 개까지 존재할 수 있다. 따라서, n개의 질의에 대한 모든 따름 관계를 일일이 화살표로 표현하려면 $O(n^2)$의 공간 복잡도가 필요하다. 대단히 많은 수의 질의들이 존재할 경우, 이들 모두를 저장하려면 저장 공간의 비용이 크다.

세 질의 Q_1, Q_2, Q_3에 대하여, $R(Q_1, Q_2, BUY)$, $R(Q_2, Q_3, BUY)$, 그리고 $R(Q_1, Q_3, BUY)$가 존재하는 경우(그림 3(a) 참고), 앞의 두 따름 관계가 있다면 $R(Q_1, Q_3, BUY)$를 삭제해도 세 질의를 따름 관계를 이용하여 처리하는 데에는 아무 문제가 없다(그림 3(b) 참고).

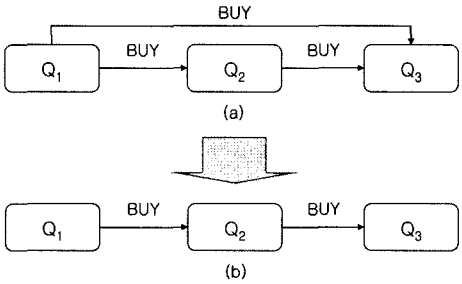


그림 3 따름 관계를 삭제해도 무방한 경우

Q를 선행 질의로 하는 모든 따름 관계 $R(Q, Q_j, X_k)$ 중, 삭제해도 무방한 따름 관계들을 모두 삭제하고 남은 따름 관계들의 집합을 아래와 같이 정의한다.

[정의 4] 최소 따름 관계 집합 mRS(Q, X)
질의 Q, 추천값 X에 대하여,
$QS_r(Q, X)$ = { Q_r | Q_r 은 R(Q, Q_r, X)를 만족하는 모든 질의 }
라 하면, 최소 따름 관계 집합 mRS(Q, X)는 다음과 같이 정의된다.

$$mRS(Q,X) = RS(Q,X) - \bigcup_{\forall Q_r} RS(Q_r,X),$$

여기서 $Q_r \in QS_r(Q,X)$ □

그림 3에 나타난 바와 같이, $R(Q_1, Q_2, X)$가 존재할 경우, 항상 $RS(Q_1, X) \supset RS(Q_2, X)$이 성립한다. 이 성질을 이용하여, 다음과 같은 간단한 알고리즘으로 mRS(Q, X)를 구할 수 있다.

[알고리즘 1] 최소 따름 관계 집합 mRS(Q, X)를 구하는 알고리즘

```
Algorithm get_mRS(Q, X)
  Input : 질의 Q, 추천값 X
  Output : 최소 따름 관계 집합 mRS(Q, X)

  Begin
    mRS(Q, X) ← RS(Q, X)
    Qr ∈ QSr(Q, X) 인 모든 Qr 에 대하여
      mRS(Q, X) ← mRS(Q, X) - RS(Qr, X)
    return mRS(Q, X)
  End
```

투자자가 새 질의 Q를 입력하면, 모든 추천값 X에 대하여 RS(Q, X)를 먼저 구한 후, 위의 알고리즘을 사용하여 mRS(Q, X)만 남긴다. 그 다음, mRS(Q, X) 안의 임의의 원소 Q_r에 대하여, $Q_r \in mRS(Q_k, X)$인 모든 질의 Q_k를 찾아서, $R(Q_k, Q_r, X)$를 $R(Q_k, Q, X)$로 치환하여 기존의 따름 관계 속에 새 질의 Q를 연결한다.

n개의 질의가 존재할 때, 새 질의 Q를 입력하는 데에 드는 시간 비용을 알아보자. $QS_r(Q, X)$ 안에 포함될 수 있는 질의 Q_r은 최대 n개까지 존재할 수 있으며, $RS(Q_r, X)$의 개수는 Q_r을 제외하면 최대 n-1개까지 존재할 수 있으므로, 총 시간은 $O(n(n-1)) = O(n^2)$이 된다. 그러나 새 질의를 입력할 때에 새 질의 Q와 종목, 시간 간격, 규칙 바디의 길이가 모두 같은 질의들만을 대상으로 하므로, 실제 소요되는 시간 비용은 크지 않다.

따름 관계를 표현하는 자료 구조를 위한 메모리 저장 공간 비용을 살펴본다. 모든 질의가 위와 같이 최소 따름 관계 집합만 유지한다고 할 때, 임의의 질의 Q가 가질 수 있는 따름 관계의 최대 개수를 생각해 보자. 예를 들어 F(Q)=SELL 일 경우, mC값은 같으나 α값이 더 큰 추종 질의와, 반대로 α값은 같으나 mC값이 더 작은 추종 질의가 존재할 수 있다. F(Q)=BUY인 경우도 마찬가지로 추종 질의는 최대 2개 존재하며, F(Q)=HOLD인 경우에는 mC, α, β 중 두 값이 같고 나머지 하나가 다른 질의들이 존재할 수 있으므로 추종 질의의 최대 개수는 3이다. 따라서, 임의의 질의 Q를 선행 질의로 하는 추종 질의의 개수는 최대 7개까지 존재할 수 있다. 따라서, 최악의 경우에도 공간 복잡도는 O(7n) = O(n)이 된다.

4.3 질의 처리 순서의 결정

모든 질의는 직접 처리되거나 혹은 간접 처리된다. 질의가 직접 처리된다는 것은 직접 질의 처리 과정을 거쳐서 추천값을 계산한다는 의미이며, 질의가 간접 처리된다는 것은 따름 관계에 의하여 선행 질의의 추천값을 그대로 적용한다는 의미이다. 투자자가 입력한 모든 질의들과 그들간의 따름 관계는 유향 그래프의 집합으로 표현된다. 이렇게 저장된 질의 집합을 대상으로, 질의 처리 비용이 최소가 되도록 질의 처리 순서를 결정해야 한다.

따름 관계 집합 RS(Q, X)의 크기가 가장 큰 질의 Q를 선택하여 먼저 처리할 때에, 간접 처리되는 질의의 수가 많아질 가능성이 높다. 그러나 RS(Q, X)에 속한 질의가 아무리 많다고 하더라도, F(Q)의 값이 X가 아닌 경우에는 간접 처리가 불가능하다. 또한, 비용 문제 때문에 모든 Q, X에 대하여 RS(Q, X)를 유지하는 것은 실질적으로 어렵다. 이러한 두 가지 이유로 인하여 본 연구에서는 따름 관계 집합의 크기가 큰 순서대로 질의들을 처리하지 않는다.

질의 처리 순서의 결정을 위하여, 따름 관계 집합 RS(Q, X)의 원소 수 대신 사용할 수 있는 기준들 중 하나는 질의의 최소 신뢰도이다. 두 질의 사이의 따름 관계가 성립하는지를 결정하는 변수는 질의의 보유 변동률 [α, β]와 최소 신뢰도 mC인데, 이들 중 추천값과 상관없이 따름 관계의 순서를 결정하는 변수는 최소 신

뢰도뿐이기 때문이다. 즉, 최소 신뢰도가 큰 질의는 추천값과 상관없이 최소 신뢰도가 더 작은 다른 질의의 선행 질의가 될 수 있으나, 그 반대의 경우는 존재할 수 없다. 따라서, 최소 신뢰도가 높은 질의일수록 따름 관계 집합의 크기도 커질 가능성이 높다는 추론이 가능하다. 이러한 고찰을 바탕으로 본 연구에서는 최소 신뢰도가 높은 질의를 먼저 처리함으로써 따름 관계를 이용한 질의의 간접 처리 가능성을 높이는 방식을 채택한다.

다음 알고리즘 2는 최소 신뢰도가 높은 순서대로 질의를 선택하여 처리하는 알고리즘이다.

[알고리즘 2] 최소 신뢰도가 높은 순서대로 질의를 선택하여 처리하는 알고리즘

```
Algorithm selectQuery(QS)
  Input : I, T, BL 값이 같은 모든 질의들의 집합 QS
  Output : 없음

  Begin
    QS에 남은 원소가 없을 때까지 반복:
      Q ← QS의 원소 중 최소 신뢰도가 가장 큰 질의
      QS에서 Q를 제거한다.
      If F(Q) 의 값이 아직 결정되지 않았다면
        F(Q)를 직접 계산한다.
      X ← F(Q)
      mRS(Q, X)에 속한 모든 질의 Q'에 대하여, F(Q')
      값을 X로 결정한다.
  End
```

QS에 새 질의를 입력했을 때, 이미 정렬된 리스트에 새 질의를 삽입하는 비용은 O(log n) 시간이 소요되므로, 4.2절에서 언급한 질의 집합 구성 시간 비용인 $O(n^2)$에는 별 영향을 주지 않는다.

5. 성능 평가

이 장에서는 실험에 의한 성능 평가를 통하여 제안하는 기법의 우수성을 규명한다.

5.1 실험 환경

본 연구에서는 실제 주가 데이타를 사용하여 성능을 측정하였다. 주가 데이타는 한국 KOSPI[14] 905종목의 분당 주가 변화 3개월 분량을 기록한 주가 데이타를 사용하였다. 실험을 위한 기본 질의 집합은 모든 종목에 대하여 다음 인수들을 사용하여 생성하였다.

- α: -0.003, -0.002, -0.001 중 한 값 선택
- β: 0.001, 0.002, 0.003 중 한 값 선택
- T: 0 으로 고정
- BL:1, 3, 5 중 한 값 선택
- mC: 50%, 60%, 70%, 80% 중 한 값 선택

위의 변수들을 사용하면, 905개 종목에 대하여 종목 하나당 각각 108(=3×3×3×4)개씩, 총97,740(=108×905)개의 질의들이 생성된다. 실험은 인텔 펜티엄4 2.4GHz, 1GB 메모리, 80GB HDD, 윈도우 2003 서버 운영체제를 사용하는 컴퓨터를 사용하여 수행하였다.

관련 연구 [10]에서는, 제안한 모델의 적중률을 검증하기 위하여 만족율과 추천율이라는 두 가지 평가 기준으로 다양한 독립 변수에 대한 실험을 수행하였다. 본 연구에서는 질의간의 따름 관계 적용에 따른 전체 질의 처리 속도 향상 효과의 검증을 위한 실험을 정의하였다. 따라서 질의 처리 속도의 향상과 관계없는 T, BL 등의 변수들에 대한 실험은 고려하지 않기로 한다.

5.2 실험 결과

질의들간의 따름 관계를 이용하여 불필요한 질의 처리 과정을 제거했을 때에, 전체 성능이 얼마나 개선되었는지를 제시한다. 질의 사이의 따름 관계를 적용할 경우, 모든 따름 관계 정보와 최소 신뢰도를 기준으로 질의들을 정렬한 리스트를 모두 메모리에 저장하기 때문에 메모리 사용량이 증가하게 된다. 전체 97,740 개의 질의에 대하여 이러한 정보를 모두 메모리에 저장한 후 측정한 결과, 전체 저장 공간의 크기는 약 12MB 정도였다. 이는 최근의 컴퓨터 메모리 사양과 비교할 때 성능 향상의 정도가 크다면 부담할 수 있는 저장 공간 비용이다. 따름 관계를 사용하지 않았을 때에는 전체 질의를 모두 처리하는 시간을 측정하였으며, 따름 관계를 사용했을 때에는 질의의 실행 순서를 결정하는 시간과 실제 질의 처리 시간을 모두 합쳐서 측정하였다.

5.2.1 따름 관계 사용에 따른 성능 개선 효과

표 1은 따름 관계를 사용하지 않고 모든 질의를 처리했을 때와 따름 관계를 사용하여 모든 질의를 처리했을 때의 전체 질의 처리 시간, 직접 처리한 질의의 수, 그리고 디스크 접근량을 비교한 결과이다.

질의간 따름 관계를 이용하여 모든 질의를 처리한 경우, 전체 질의 처리 시간이 9.5% 로 감소하여 처리 속도가 비약적으로 빨라졌음을 확인할 수 있다. 실제로 질의

표 1 따름 관계 적용시와 미적용시의 질의 수행 시간, 디스크 접근량, 질의수와 그 비

	전체 질의 처리 시간 (초)	직접 처리한 질의 수(개)	디스크 접근량 (byte)
따름 관계 미적용시	3,553.01	1,509,300	179,567,058,384
따름 관계 적용시	338.40	288,024	18,457,371,156
감소 비율 (%)	90.5	81	89.7

를 수행한 회수를 비교하면, 약 150만의 전체 질의들 중 약 122만의 질의가 직접 처리되지 않고 따름 관계에 의하여 간접적으로 그 추천값이 결정되었음을 알 수 있다.

그런데, 질의를 모두 직접 처리하지 않고 따름 관계를 이용하는 경우, 따름 관계를 적용하기 위한 추가 시간이 더 소요된다. 따라서 직접 처리되지 않은 비율인 81%보다 처리 시간이 덜 줄어들 것이라고 예상할 수 있다. 그러나 실제 실험 결과에서는 전체 질의 처리 시간이 90% 이상 줄어들었음을 확인할 수 있다. 이러한 결과의 원인으로 두 가지를 들 수 있다. 첫째, 각 질의마다 처리 시간이 다르기 때문이며, 둘째, 디스크 접근 시간이 질의 처리 시간의 가장 큰 비중을 차지하기 때문이다. 실험 결과에서 디스크 접근량의 감소 비율과 전체 질의 처리 시간의 감소 비율이 거의 같은 것은 이러한 원인을 뒷받침하는 증거이다.

5.2.2 질의의 개수 증가에 따른 성능 개선 효과

그림 4는 전체 질의의 개수를 2, 4, 8배로 증가시켜 가며 전체 질의를 모두 처리하는 데 걸린 시간을 측정한 결과를 나타낸 것이다.

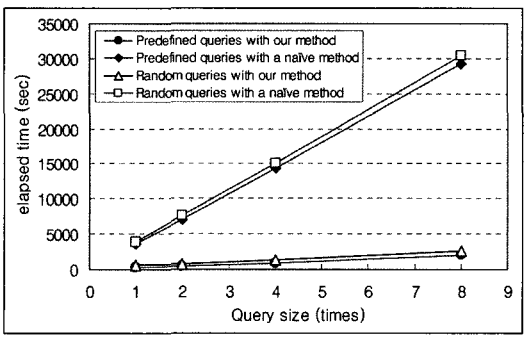


그림 4 질의 개수 증가에 따른 질의 처리 시간

그림 4에서 predefined query는 5.1절에서 정의하고 5.2절의 실험에서 사용한 기본 질의 집합이며, random query는 5.1절의 각각의 변수에 대하여 무작위로 결정한 값들을 사용하여 만든 질의 집합을 가리킨다. 무작위로 변수의 값들을 결정하여 질의를 구성할 경우, 실험에서 사용한 기본 질의 집합에 비해 질의 사이의 따름 관계의 발생 가능성이 줄어든다. 두 경우의 전체 질의 개수는 서로 같도록 하였다.

전체 질의가 모두 질의 사이의 따름 관계의 적용을 받는 predefined query에 비하여, 그렇지 않은 random query의 수행 시간이 조금 더 긴 결과를 보이고 있으나, 그 폭은 크지 않다. 이것은 random query의 경우 따름 관계를 적용받지 않는 질의들이 생기기 때문이며, 그 수만큼 실제 질의 수행의 회수가 증가하기 때문이다. 그러나, 두 실험 시간에는 큰 차이가 없는 것으로 나타났다. 이는 실험에서 사용한 질의 집합이 무작위로 생성된 질의 집합에 비해 특별히 제안된 기법의 우수성을 보이기 위하여 인위적으로 구성한 질의가 아님을 보이는 것이다.

실험 결과를 보면, 질의 사이의 따름 관계를 사용하지 않았을 경우에는 모든 질의를 개별적으로 처리해야 하므로, 질의 개수의 증가에 비례하여 질의 처리 시간이 길어지게 된다. 이 경우에는 predefined query와 random query 사이의 처리 시간이 별 차이가 없는데, 이는 처리하는 질의의 개수가 거의 동일하기 때문이다. 한편, 질의 사이의 따름 관계를 적용하여 질의를 처리할 경우, 전체 수행 시간이 비약적으로 감소하며, 그 감소 폭은 질의의 개수가 증가할수록 커지는 좋은 특성을 보인다. 제안된 기법은 기존 기법과 비교하여 최대 14배까지의 성능 개선 효과를 보이는 것으로 나타났다.

5.2.3 질의의 최소 신뢰도와 따름 관계 집합의 크기 사이의 관계

4.3절에서는 최소 신뢰도가 높은 질의일수록 따름 관계 집합의 크기도 커질 가능성이 높다는 추론을 바탕으로 질의의 최소 신뢰도를 질의 처리 순서를 정하는 기준으로 사용하는 알고리즘을 제안하였다. 그림 5는 무작위로 결정한 5.1절의 각각의 변수 값들을 사용하여 만든 97,740개의 질의 집합과 모든 추천값 X∈{BUY, HOLD, SELL}에 대하여, 각 질의의 최소 신뢰도와 따름 관계 집합 RS(Q, X)에 속하는 질의들의 평균 수를 비교한 결과이다.

질의 Q의 최소 신뢰도가 높을수록 따름 관계 집합 RS(Q, X)에 속하는 질의의 평균 수가 많아짐을 알 수 있다. 두 값 사이의 상관 계수(correlation)는 0.99로, 통계적으로도 대단히 강한 양의 상관관계(positive correlation)를 가진다. 따라서 따름 관계 집합에 속하는 질의의 수 대신 질의의 최소 신뢰도 순서대로 질의를 처리해도 거의 같은 효과를 얻는다는 사실을 알 수 있다.

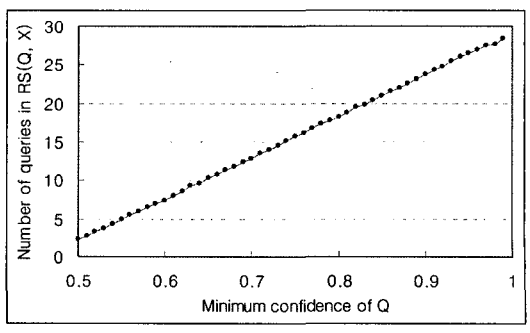


그림 5 질의의 최소 신뢰도에 따른 따름 관계 집합의 크기

6. 결 론

본 논문에서는 다수의 사용자로부터 입력된 전체 질의를 연속적으로 처리해야 하는 환경에서 효과적으로 질의를 처리하는 방안에 관하여 논의하였다.

이전 연구에서 제안된 방식은 주식 데이타베이스로부터 탐사된 규칙들을 대상으로 하는 투자자의 질의를 통하여 향후 주식의 가격 추이를 예측하고, 이를 기반으로 투자자에게 주식의 보유, 매수, 매도 등의 투자 행위를 추천한다. 이를 위하여 새로운 규칙 모델을 정의하고, 빈번하게 발생하는 주가 변화 패턴의 이후의 경향이 투자자의 투자 조건과 매치하는 경우 이를 규칙으로 생성한다. 이때, 빈번하게 발생하는 패턴을 규칙의 헤드로 정의하고, 이후의 주가 변화 경향을 규칙의 바디로 정의한 후 이를 규칙 모델로 사용하였다.

본 연구에서는 투자자들이 입력한 대용량의 질의들이 연속으로 반복 실행되는 환경에서 전체 질의를 효과적으로 처리하기 위한 새로운 기법을 제안하였다. 투자자가 입력하는 질의들의 특성을 분석하여 추천값이 자동으로 결정될 수 있는 따름 관계를 정의하고, 실제로 따름 관계를 가지는 질의들을 찾아서 중복 실행되는 질의들을 직접 실행하지 않도록 질의 처리 순서를 결정한다. 이 결과, 전체 질의 처리 시간을 크게 개선할 수 있다. 이와 같은 작업을 효율적으로 수행하기 위하여, 따름 관계를 가지는 질의들을 찾은 뒤 최소 따름 관계 집합만을 구성하는 알고리즘과, 구성된 따름 관계가 효율적으로 적용되도록 질의 처리 순서를 결정하는 알고리즘을 제안하였다.

제안한 방법의 우수성을 규명하기 위하여 따름 관계를 사용하는 기법과 그렇지 않은 기법에 대한 실험을 수행하였다. 실험 결과에 의하면 제안된 질의간 따름 관계를 적용하는 알고리즘을 사용하였을 경우, 이전에 비하여 90% 정도의 질의 처리 시간 감소 효과를 볼 수 있었다.

향후에는 지속적으로 입력되는 주가 데이타를 빈번 발생 패턴 집합에 동적으로 반영하는 기법[15]에 대하여 연구하고자 한다.

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] 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. Int'l. Conf. on Data Engineering, IEEE, pp. 607-614, 2001.

[3] W. K. Loh, S. W. Kim, and K. Y. Whang, "A Subsequence Matching Algorithm that Supports Normalization Transform in Time-Series Databases," Data Mining and Knowledge Discovery Journal, Vol. 9, No. 1, pp. 5-28, July. 2004.

[4] S. H. Park et al., "Efficient Searches for Similar Subsequences of Difference Lengths in Sequence Databases," In Proc. Int'l. Conf. on Data Engineering, IEEE ICDE, pp. 23-32, 2000.

[5] P. Bloomfield, Fourier Analysis of Time Series, Wiley, 2000.

[6] R. Agrawal et al., "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases," In Proc. Int'l. Conf. on Very Large Data Bases, VLDB, pp. 490-501, Sept. 1995.

[7] 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.

[8] T. Anderson, The Statistical Analysis of Time Series, Wiley, 1971.

[9] T. Hellstrom and K. Holmstrom, Predicting the Stock Market, Opuscula ISRN HEV-BIB-OP-26-SE, Aug. 1998.

[10] Authors removed, "Rule Discovery and Matching in Stock Databases," submitted for publication, 2006.

[11] S. Babu and J. Widom, "Continuous Queries over Data Streams," ACM SIGMOD Record Vol. 30, No. 3, pp. 109-120, 2001.

[12] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," In Proc. Int'l. Conf. on Very Large Data Bases, VLDB, pp. 487-499, 1994.

[13] R. Agrawal and R. Srikant, "Mining Sequential Patterns," In Proc. Int'l. Conf. on Data Engineering, IEEE ICDE, pp. 3-14, 1995.

[14] Koscom Data Mall, http://datamall.koscom.co.kr, 2005.

[15] G. Manku and R. Motwani, "Approximate Frequency Counts over Data Streams," In Proc. Int'l. Conf. on Very Large Data Bases, VLDB, pp. 346-357, Aug. 2002.

하 유 민
2005년 2월 연세대학교 컴퓨터과학과 졸업(학사). 2005년 3월~현재 연세대학교 컴퓨터과학과 석사과정. 관심분야는 데이타 마이닝, 스트림 데이타베이스, 멀티미디어 데이타베이스 등

김 상 욱
정보과학회논문지 : 데이타베이스
제 33 권 제 2 호 참조

박 상 현
정보과학회논문지 : 데이타베이스
제 33 권 제 3 호 참조