주가 예측을 위한 규칙 탐사 및 매칭
(Rule Discovery and Matching for Forecasting Stock Prices)

하 유 민 † 김 상 욱 †† 원 정 임 ††† 박 상 현 †††† 윤 지 희 †††††
(You-Min Ha) (Sang-Wook Kim) (Jung-Im Won) (Sang-Hyun Park) (Jee-Hee Yoon)

요 약 본 논문에서는 주식 데이타베이스로부터 과거 주가 변화 패턴에 대한 규칙을 탐사함으로써 투자자에게 주식 투자 유형을 추천해 주는 방안에 관하여 논의한다. 먼저, 본 논문에서는 주식 투자 유형의 추천을 위한 새로운 규칙 모델을 정의한다. 제안된 모델에서는 빈번하게 발생하는 주가 변화 패턴의 이후의 주가 변화 경향이 투자자의 투자 조건과 매치하는 경우, 이 종목에 대한 투자 유형을 추천하도록 하는 방식을 사용한다. 이때, 빈번하게 발생하는 패턴을 규칙의 헤드로 간주하며, 이후의 주가 변화 경향을 규칙의 바디로 간주한다. 본 연구에서는 규칙 헤드는 투자자의 특성에 별다른 영향을 받지 않는 반면, 규칙 바디에 대한 조건은 투자자마다 다르다는 점에 착안하여 규칙 탐사 과정에서 전체 규칙이 아닌 규칙 헤드들만을 탐사하여 저장해 두는 새로운 방식을 제안한다. 이 결과, 투자자 별로 달라질 수 있는 규칙 바디에 대한 조건을 유연하게 정의하는 것을 허용하며, 규칙의 수를 줄임으로써 전체 규칙 탐사 성능을 개선할 수 있다. 효율적인 규칙 탐사와 매칭을 위하여 빈번 패턴들을 효과적으로 탐사하는 방법, 빈번 패턴 베이스를 구축하는 방법, 그리고 이들을 인덱싱 하는 방법을 제안한다. 또한, 투자자의 질의가 발생하는 경우, 빈번 패턴 베이스로부터 이와 매치되는 규칙을 발견하고, 이 결과를 이용하여 투자자에게 투자 유형을 추천해 주는 방법을 제안한다. 실제 주식 데이타를 이용한 다양한 실험을 통하여 제안된 기법의 우수성을 규명한다.

키워드 : 주식 데이타베이스, 규칙 탐사, 규칙 매칭

Abstract This paper addresses an approach that recommends investment types for stock investors by discovering useful rules from past changing patterns of stock prices in databases. First, we define a new rule model for recommending stock investment types. For a frequent pattern of stock prices, if its subsequent stock prices are matched to a condition of an investor, the model recommends a corresponding investment type for this stock. The frequent pattern is regarded as a rule head, and the subsequent part a rule body. We observed that the conditions on rule bodies are quite different depending on dispositions of investors while rule heads are independent of characteristics of investors in most cases. With this observation, we propose a new method that discovers and stores only the rule heads rather than the whole rules in a rule discovery process. This allows investors to define various conditions on rule bodies flexibly, and also improves the performance of a rule discovery process by reducing the number of rules. For efficient discovery and matching of rules, we propose methods for discovering frequent patterns, constructing a frequent pattern base, and indexing them. We also suggest a method that finds the rules matched to a query issued by an investor from a frequent pattern base, and a method that recommends an investment type using the rules. Finally, we verify the superiority of our approach via various experiments using real-life stock data.

Key words : stock databases, rule discovery, rule matching

· 본 논문은 제주대학교를 통한 정보통신부 및 정보통신진흥원의 대한 IT연구센터지원사업(IITA-2005-C1090-0502-0009)의 연구비 지원을 받았습니다.

† 학생회원 : 연세대학교 컴퓨터과학과
ymha@cs.yonsei.ac.kr
†† 종신회원 : 한양대학교 정보통신학부 교수
wook@hanyang.ac.kr
††† 정 회 원 : 한양대학교 정보통신학부 교수
jiwon@hanyang.ac.kr
†††† 종신회원 : 연세대학교 컴퓨터과학과 교수
sanghyun@cs.yonsei.ac.kr
††††† 종신회원 : 한림대학교 정보통신학부 교수
jhyoon@hallym.ac.kr
논문접수 : 2005년 9월 16일
심사완료 : 2007년 3월 30일

1. 서 론

시계열 데이타(time-series data)란 시간의 흐름에 따라 일정 간격으로 객체의 변화를 관측하여 얻어진 값들의 리스트이다[1-4]. 이러한 시계열 데이타는 경제 현상이나 자연 현상 등에 관한 시간적 변화를 나타내는 데이타이며, 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 값

을 예측할 수 있다. 주가의 변화를 기록한 데이타는 대표적인 시계열 데이타의 하나이다[6,7]. 주식 투자자의 궁극적인 목적은 수익률을 극대화하는 것이므로, 주식 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.

주식 데이타의 미래 예측을 위하여 시계열 분석(time-series analysis) 기법이 널리 사용되고 있다. 시계열 분석은 크게 시간 영역 분석과 주파수 영역 분석으로 분류된다. 시간 영역 분석은 현재 시점의 값이 과거 시점의 값들에 의해 결정되는 회귀 모델을 기반으로 하는 방식이다[8]. 주파수 영역 분석은 정적인 시계열 데이타(stationary time-series data)를 분석할 때에 주로 사용되며, 월별, 계절별, 년도별 등의 거시적인 변화 추세를 예측하는 데에 유용하다[5]. 그러나 이러한 기법들은 주식의 변화 예측을 위한 투자자의 동적인 요구를 반영할 수 없으며, 단시간 내의 주가 변화 예측에는 적용하기 어렵다는 문제점이 있다.

한편, 기계 학습(machine learning) 분야에서도 신경망(neural network)을 이용하여 과거 주식 데이타를 분석함으로써 미래의 주가를 예측하는 기법들이 있다[9-11]. 그러나 이러한 기법들에서 사용자의 동적인 요구를 수용하기 위해서는 사용자가 지정한 각 요구 조건에 대해 신경망을 개별적으로 구성해야 하므로 주기억 장치의 큰 낭비를 초래한다. 또한, 이러한 기법들은 대용량의 데이타베이스 환경에 적용하기에는 성능상의 문제점이 있다.

시계열 데이타베이스로부터 규칙을 탐사하고 매칭하는 문제를 데이타베이스 관점에서 다룬 연구들도 있었다[12,13]. 참고 문헌 [12]에서는 시계열 데이타를 심볼 시퀀스로 변환한 후, 이로부터 규칙을 탐사하는 기법을 제안하였다. 심볼 시퀀스로의 변환을 위하여 시계열 데이타로부터 다수의 윈도우(window)들을 추출한 후, 전체 윈도우들을 클러스터링 방법[14]을 이용하여 그루핑한다. 그 다음, 각 윈도우 그룹에 고유의 심볼을 대응시켜, 전체 시계열 데이타를 심볼 시퀀스로 변환한다. 참고 문헌 [13]에서는 탄력 규칙(elastic rule)이라는 개념을 도입하고, 접미어 트리(suffix tree)를 이용하여 탄력 규칙들을 탐사하는 기법을 제안하였다. 이때, 시계열 데이타를 심볼 시퀀스로 변화하기 위한 방법으로서 TAH 트리(TAH tree)[15]를 이용하였다.

규칙은 규칙 헤드(rule head)와 규칙 바디(rule body)로 구성된다. 데이타베이스 분야에서 수행된 기존의 기법들이 가지는 공통점은 규칙 헤드와 바디를 규칙 탐사 과정에서 모두 찾아낸다는 점이다. 이것은 규칙 헤드 및 바디에 해당되는 조건을 규칙 탐사 이전에 미리 정의해야 한다는 것을 의미한다. 그러나 제 2장에서 자세히 언급되는 바와 같이 주식 투자 유형의 추천에서 사용되는 규칙 바디에 대한 조건은 투자자의 성향에 따라 다르다. 따라서 이러한 상황에서 규칙 헤드와 바디를 함께 탐사하는 기존 기법들은 지나치게 많은 규칙들을 생성시키게 되며, 아울러 새로운 조건을 기존 탐사 결과에 반영할 수 없다는 두 가지 문제점들을 갖는다.

본 논문에서는 이러한 문제점들을 해결할 수 있는 새로운 규칙 탐사 및 매칭 기법을 제안하고, 이를 주식 투자에 응용하는 방안에 관하여 논의한다. 규칙 헤드와 바디를 모두 탐사하는 기존의 기법들과는 달리, 제안된 기법에서는 규칙 헤드들만을 탐사하여 저장해 두는 방식을 제안한다. 투자자가 관심 종목에 대하여 질의하면, 이와 매치되는 규칙 헤드를 찾아내고, 규칙 바디에 해당되는 부분은 이 시점에 탐사한다. 따라서 제안된 기법은 투자자들이 규칙 바디의 조건을 다양하게 정의하는 것을 허용하여, 여러 투자자들의 다양한 질의를 모두 처리할 수 있으며, 단기간의 주가 예측 질의가 입력되어도 처리할 수 있다. 또한, 투자자들이 입력한 질의 데이타와 규칙 처리를 위한 패턴 데이타를 따로 저장하여 처리함으로써, 사용되는 저장 공간을 줄일 수 있다.

본 논문의 주요 공헌은 아래와 같다.

- 주식 투자 유형의 추천을 위한 유연성 있는 새로운 규칙 모델을 정의한다.
- 주식 데이타베이스로부터 규칙의 헤드에 해당되는 빈번 패턴(frequent pattern)[16]들을 효과적으로 탐사하는 방법을 제안한다.
- 탐사된 빈번 패턴들을 빈번 패턴 베이스의 형태로 구축하고, 이들을 효과적으로 인덱싱 하는 방법을 제안한다.
- 빈번 패턴 베이스를 이용하여 효과적으로 규칙을 매칭하고, 이 결과를 이용하여 투자자에게 투자 유형을 추천해 주는 방법을 제안한다.
- 제안된 기법의 우수성을 규명하기 위하여 다양한 실험을 통한 성능 평가를 수행한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 논문에서 해결하고자 하는 문제를 정의하고, 이를 해결하는 제안된 기법의 전체적인 개관을 설명한다. 제3장에서는 빈번 패턴 베이스를 구축하는 방법과 이를 인덱싱하는 방법을 차례로 설명한다. 제4장에서는 규칙 매칭 방법과 이를 이용하여 투자 유형을 추천하는 방법에 관하여 기술한다. 제5장에서는 다양한 실험에 의한 성능 평가를 통하여 제안된 기법의 우수성을 규명한다. 끝으로, 제6

장에서는 본 논문을 요약하고, 결론을 내린다.

2. 개 관

본 장에서는 제안하는 기법에 관하여 간략히 기술한다. 먼저, 제2.1절에서는 본 연구의 동기가 되는 응용 예제를 제시하고, 제2.2절에서는 제안하는 기법에서 채택하는 규칙 모델에 관하여 설명한다. 제2.3절에서는 제안하는 기법이 가지는 주요 특성에 대하여 논의한다.

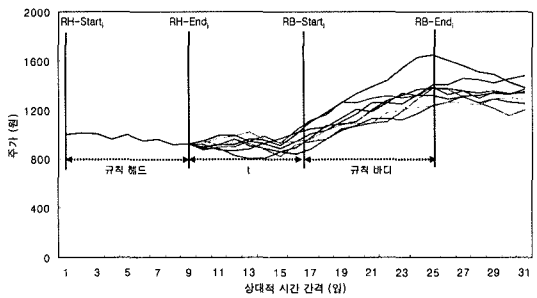

그림 1 직관적 예제

그림 1은 한 주식 종목의 열 개의 주가 변화 패턴 Stocki (0<=i<10)을 도면화한 것이다. 세로축은 주가를 나타내며, 가로축은 해당 주가 패턴이 출현한 상대 시간 구간을 나타낸다. RH-start$_i$와 RH-end$_i$는 각각 '규칙 헤드'라 표기된 시간 구간 내의 주가 패턴 Stock$_i$의 첫 값과 끝 값이 출현한 시각을 의미한다. 유사한 방식으로, RB-start$_i$와 RB-end$_i$는 각각 '규칙 바디'라 표기된 시간 구간 내의 주가 패턴 Stock$_i$의 첫 값과 끝 값이 출현한 시각을 의미한다.

그림의 규칙 헤드 구간을 보면 열 개의 주가의 변화 패턴들은 완전히 일치하고 있다는 것을 알 수 있다. 이 때, 규칙 헤드 구간으로부터 일정한 시간 간격 t 후에 나타나는 규칙 바디 구간 안의 주가의 변화 패턴을 살펴보자. 이 구간 내에서, 모든 패턴들의 평균 주가는 규칙 헤드 구간의 마지막 주가와 비교하여 상승하였음을 보이고 있다. 이로 미루어 규칙 헤드 구간과 같은 주가 변화 패턴을 보이는 종목들은 t 시간 후에는 그 주가가 상승하였다고 해석할 수 있다.

이러한 주가 변화의 경향을 어떤 투자자가 미리 알고 있다고 하자. 만일, 이 투자자가 관심을 가지던 종목의 주가가 규칙 헤드 구간의 패턴을 보이는 순간 이 투자자는 이 종목의 주가가 시간 t 이후 강하게 상승한다고 예측할 수 있다. 이 투자자는 이러한 종목을 매수함으로써 높은 수익을 올릴 수 있는 기회를 얻을 수 있을 것이다. 본 연구의 목적은 이러한 과거의 주가 변화 패턴을 분석함으로써 투자자들에게 주식 투자 유형을 자동적으로 추천해 주는 시스템을 개발하는 것이다. 다음 절에서 이러한 모델을 명확히 정의하고, 3장과 4장에서는 실제로 작동하는 구체적인 알고리즘을 설명한다.

2.2 규칙 모델

본 논문에서는 제 2.1절의 직관적 예제에 나타난 주가 변화의 경향을 아래와 같은 형태의 규칙을 사용하여 표현한다.

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

여기서, H는 규칙 헤드(rule head)라 하며, B는 규칙 바디(rule body)라 부른다. 이 규칙은 H에 해당되는 사건이 발생한 후, t 시간이 흐른 후에는 B에 해당되는 사건이 발생하였음을 의미한다.

본 응용에서 H는 그림 1에 나타난 예제의 'rule head' 구간에서와 같은 특정 주가 변화 패턴 P의 발생과 대응되는 사건이다. 또한, B는 'rule body' 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 'rule body' 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 'rule head' 구간의 마지막 주가 대비 'rule body' 구간에서의 평균 주가 상승률이 20% 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.

다음은 (s, c)에 대하여 설명한다. 주가 변화의 패턴이 규칙으로서 가치를 가지기 위해서는 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.

$$support(H) = \frac{H\text{와 매치하는 패턴들의 발생수}}{H\text{와 매치하는 패턴과 길이가 동일한 패턴들의 전체 발생수}} \times 100$$

또한, 규칙으로서 가치를 가지기 위해서는 H와 매치하는 과거 패턴들이 'rule body' 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.

$$\text{confidence}(H,B) = \frac{H\text{와 매치하며, } B\text{의 조건을 만족시키는 패턴들의 발생 수}}{H\text{와 매치하는 패턴들의 발생 수}} \times 100$$

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.

2.3 제안된 기법의 주요 특징

규칙 탐사에 관한 기존의 연구에서는 규칙 탐사 과정에서 헤드와 바디를 모두 포함하는 규칙 들을 찾아내는 방식을 사용한다. 따라서 규칙 바디 B를 위한 조건의 정의를 규칙 탐사 이전에 미리 지정해야 한다. 그러나 이러한 조건에 대한 정의는 투자자마다 달라질 수 있다. 예를 들면, 어떤 투자자는 규칙 헤드의 ***마지막 주가 대비*** 규칙 바디의 ***평균*** 상승률이 20% 이상인 규칙에 의하여 매수 추천을 받기를 원하는 반면, 또 다른 투자자는 규칙 헤드의 ***평균 주가 대비*** 규칙 바디의 ***최저*** 상승률이 10% 이상인 규칙에 의하여 매수 추천을 받기를 원할 수 있기 때문이다. 뿐만 아니라, 이러한 투자 유형 추천에 대한 기준은 동일한 투자자라 할지라도 투자 시점에 따라 달라질 수 있다.

이러한 상황에서 규칙 헤드와 바디를 함께 탐사하는 기존 방식을 사용하는 경우, 다음과 같은 문제점들이 발생한다. 첫째, 규칙 탐사 시점에 각각의 투자자가 정의한 규칙 바디 조건을 만족하는 모든 규칙들을 탐사해야 하므로 탐사해야 할 규칙들의 수가 매우 많아진다. 이 결과, 규칙 탐사 시간이 길어지며, 이후 매칭을 위하여 관리해야 할 규칙들의 저장 공간의 크기가 커진다. 둘째, 규칙 탐사 후 새로운 투자자가 원하는 규칙 바디 조건이나 기존의 투자자가 새롭게 원하는 규칙 바디 조건을 정의할 수 없다. 즉, 이러한 새로운 조건을 만족하는 규칙들은 새로운 규칙 탐사 과정을 거쳐야만 규칙 베이스에 반영될 수 있다. 따라서 규칙 탐사의 유연성이 저하된다.

또 다른 방식은 별도의 일괄적인 규칙 탐사 과정 없이 규칙 탐사와 매칭을 동일한 단계에서 수행하는 것이다. 즉, 투자자가 질의하는 시점에서 해당 관심 주식의 주가 변화 패턴과 규칙 바디 조건을 함께 입력받아 해당 종목에 대한 규칙 헤드와 바디를 즉시 탐사하여 투자 유형을 추천하는 것이다. 이러한 방식은 기존의 방식의 두 가지 문제점들을 해결할 수 있다는 긍정적인 측면이 있으나, 매 질의 시점 마다 전체 데이타베이스를 분석해야 하므로 규칙 탐사의 성능이 크게 저하된다는 문제점을 갖는다.

본 연구에서는 투자자들의 특성에 의하여 달라지는 것은 규칙 바디 B의 조건이며, 규칙 헤드H는 투자자의 특성에 크게 영향을 받지 않는 점에 착안하였다. 따라서 본 논문에서는 규칙 탐사 과정에서 전체 규칙이 아닌 규칙 헤드들만을 탐사하여 저장해 두는 방식을 제안한다. 이때, 저장된 규칙 헤드들의 집합을 빈번 패턴 베이스(frequent pattern base)라 정의한다. 투자자가 관심 종목에 대하여 질의하면, 먼저 이와 매치되는 규칙 헤드를 빈번 패턴 베이스로부터 찾고, 규칙 바디 B에 대한 투자자의 조건을 만족하는 사건의 발생 여부를 이 시점에서 확인한다. 이 결과, 제안된 기법은 투자자별로 달라질 수 있는 규칙 바디 B의 조건을 유연하게 정의하는 것을 허용하며, 규칙 탐사 성능의 저하 문제도 함께 해결할 수 있다는 큰 장점을 갖는다.

하나의 논점은 규칙들을 추출하는 대상을 무엇으로 할 것인가 하는 것이다. 한 가지 방식은 앞에서 기술한 바와 같이 전체 주식 데이타베이스를 대상으로 규칙들을 추출하는 것이고, 다른 한 방식은 각 종목 주가 시퀀스를 대상으로 종목별 규칙들을 추출하는 것이다. 전자의 경우, 전체 주가에 대한 경향을 파악할 수 있으며, 후자의 경우, 종목별 주가에 대한 경향을 파악할 수 있다. 본 논문에서 제안하는 방법은 이러한 두 가지 경우에 모두 적용이 가능하다. 제 5장에서는 종목별 주가 경향을 파악하기 위한 규칙들을 추출하는 경우에 초점을 맞추어 규칙의 만족율과 규칙 탐사 및 규칙 매칭의 성능을 분석한다.

3. 빈번 패턴 베이스의 구축

본 장에서는 주식 데이타베이스로부터 빈번 패턴 베이스를 구축하는 방안에 대하여 논의한다. 제 3.1절에서는 규칙 헤드에 해당하는 빈번 패턴들을 효과적으로 탐사하는 방법을 제안하고, 제 3.2절에서는 탐사된 빈번 패턴들을 빈번 패턴 베이스의 형태로 구축하고, 이들을 효과적으로 인덱싱 하는 방안을 설명한다.

3.1 빈번 패턴의 탐사

어떤 패턴의 지지도가 사전에 응용에서 지정한 최소 지지도(minimum support) 보다 크거나 같은 경우, 그 패턴을 빈번 패턴(frequent pattern)이라 부른다. 제 2.2절에서 설명된 바와 같이 본 연구의 규칙 모델에서는 빈번 패턴만이 규칙의 헤드로 사용될 수 있다. 그림 2는 주식 데이타베이스를 구성하고 있는 각 시퀀스로부터 빈번 패턴들을 탐사하는 전체 과정을 보여준다.

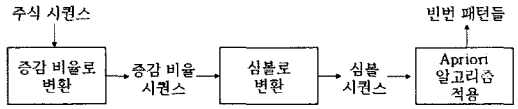

그림 2 주식 시퀀스로부터 빈번 패턴들을 탐사하는 전체 과정

3.1.1 사전 처리 과정

주식 시퀀스 내 요소 값인 주가의 크기나 변화 범위는 종목에 따라 매우 큰 차이를 보인다. 또한, 같은 종목 내에서도 전반적인 경기의 흐름이나 사업 실적 등에 따라 주가는 매우 다양한 값을 가진다. 따라서 실제 주가를 대상으로 규칙을 탐사하는 경우 더 정확한 패턴을 찾아낼 수 있으나, 최소 지지도 이상의 빈번 패턴이 존재할 가능성이 매우 낮다. 이러한 문제를 해결하기 위해서 본 연구에서는 주식 시퀀스의 각 요소 값을 다음 요소 값과의 증감 비율로 표현한 후, 이를 다시 심볼로 변환하는 방법을 사용한다. 즉, 증감 비율로 변환한 시퀀스 S′의 각 요소 값 s′[i] ($0 \le i < n-1$)는 원래의 주식 시퀀스 S의 각 요소 값 s[i] ($0 \le i < n$)들을 이용하여 다음과 같이 계산된다.

$$s'[i] = \frac{s[i+1] - s[i]}{s[i]} \times 100$$

예를 들어, 시퀀스 S = <5000, 5100, 4900, 5400>를 증감 비율로 변환하면, 시퀀스 S′ = <2.00%, -3.92%, 10.20%>가 된다. 증감 비율로 표현된 시퀀스 S′은 다시 도메인 분류 방식(categorization)에 의하여 심볼 시퀀스 S″로 변환된다. 도메인 분류는 요소 값을 심볼로 변환하기 위하여 요소 값의 도메인을 여러 범위들로 분할하는 작업이다. 서로 다른 요소 값이라도 같은 범위에 속하게 되면 동일한 심볼로 변환된다. 이 결과, 변환된 심볼 시퀀스 내에는 빈번 패턴이 존재할 가능성이 상대적으로 높아진다.

도메인 분류 방식으로 단순히 도메인을 일정 범위로 균등하게 분할하는 방식을 사용할 수 있다. 그러나 실제의 한국의 코스피(KOSPI) 주식 데이타베이스의 경우, 하루에 변화할 수 있는 주가 범위는 전날 종가의 -15%에서 +15%까지로 제한되며, 이 범위 내에서도 주가의 변화 분포가 균등한 것이 아니라 증감 비율 0%의 근처에 대부분 집중되어 있으며 -15% 또는 +15%에 가까울수록 발생 빈도는 확연히 줄어든다. 따라서 본 연구에서는 전체 도메인을 증감 비율에 따른 분포 면적이 균등하도록 분할하는 방식을 채택한다.

그림 3.2는 도메인을 9개의 범위로 분할하는 예를 보인다. 예를 들어, 실제의 주식 시퀀스 S를 증감 비율로 변환한 시퀀스 S′ = <2.00%, -3.92%, 10.20%>은 이 도메인 분류 방식에 의하여 각 요소 값이 심볼로 변환된 시퀀스 S″ = <F, C, I>이 된다. 본 연구에서는 주식 시퀀스의 각 요소 값을 심볼로 변환한 최종 시퀀스를 대상으로 빈번 패턴을 탐사한다.

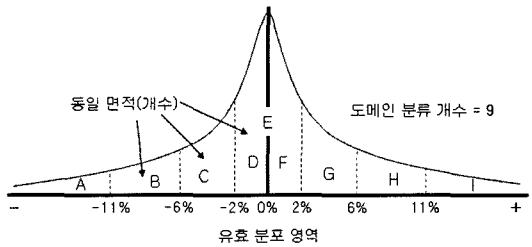

그림 3 도메인 분류 방식의 예

3.1.2 탐사 과정

데이타베이스에 저장된 각각의 심볼 시퀀스를 대상으로 가능한 모든 부분 시퀀스들을 추출한 후, 이들의 지지도를 계산해서 사전에 지정된 최소 지지도와 비교함으로써 빈번 패턴들을 탐사할 수 있다. 이 방식은 구현은 간단하지만 각 부분 시퀀스의 지지도를 계산할 때마다 전체 데이타베이스를 액세스해야 하므로 탐사 오버헤드가 매우 크다.

참고 문헌 [6,17,18]에서는 이러한 단점을 극복하기 위한 방안으로 후보 패턴의 개념을 이용한 Apriori 알고리즘을 제안하였다. 후보 패턴(candidate pattern)이란 빈번 패턴이 될 가능성이 있는 패턴을 의미한다. 예를 들어, 길이 1인 패턴 <A>, <B>, <C> 중에서 패턴 <A>, <B>는 빈번 패턴이며, 패턴 <C>는 빈번 패턴이 아니라고 가정하자. 길이 2인 빈번 패턴을 탐사하는 경우, 패턴 <AB>, <BA>는 빈번 패턴이 될 가능성이 있지만 패턴 <AC>, <BC>는 빈번 패턴이 될 가능성이 전혀 없다. 왜냐하면, 패턴 <AC>, <BC>의 부분 집합에 해당하는 패턴 <C>의 지지도가 최소 지지도보다 작기 때문이다. 즉, 모든 부분 집합이 빈번 패턴인 패턴만이 빈번 패턴이 될 가능성이 있다.

본 연구에서는 빈번 패턴을 효과적으로 탐사하기 위하여 후보 패턴 개념을 활용한 Apriori 알고리즘을 다음과 같이 사용한다. 먼저, 각 심볼 시퀀스를 대상으로 길이가 1인 빈번 패턴 집합 F_1을 생성한다. 그 다음, 길이가 k ($2 \le k \le Len(S)$)인 빈번 패턴 집합 F_k을 생성하기 위하여 길이 k-1의 빈번 패턴 집합 F_{k-1}을 셀프 조인(self-join)하여 후보 패턴 집합 C_k를 생성한다. F_k는 C_k 내의 후보 패턴 중에서 최소 지지도 이상을 만족하는 패턴만을 포함하며, 이것은 데이타베이스 전체를 한 번 스캔함으로써 파악할 수 있다. 만약, 빈번 패턴 집합 F_k가 공집합이라면 더 이상 새로운 빈번 패턴을 발견할 수 없으므로 탐사를 종료한다.

3.2 빈번 패턴 베이스의 구축

탐사된 빈번 패턴들은 빈번 패턴 베이스 형태로 저장된다. 효율적인 규칙 매칭을 위하여 빈번 패턴 베이스는 빠르게 검색할 수 있는 형태로 조직되어야 한다. 본 연구에서는 이를 위하여 B 트리를 이용하여 빈번 패턴 베이스를 구축한다. 이때, B 트리에 저장되는 엔트리는 <빈번 패턴, 발생 위치 리스트>의 쌍이다. 즉, 빈번 패턴을 표현하는 문자열이 B 트리의 키로 사용되며, 발생 위치 리스트는 해당 빈번 패턴이 발생되는 시퀀스 내의 위치들을 의미한다. 각 위치는 주식 시퀀스의 식별자(sequence identifier: SID)와 시퀀스 내에서의 오프셋(offset)으로 표현된다.

하나의 B 트리만을 이용하여 빈번 패턴 베이스를 구축할 수도 있으나, 본 연구에서는 보다 효율적인 규칙 매칭을 지원하기 위하여 빈번 패턴들을 길이를 기준으로 하여 다수의 그룹으로 분류한 후, 각 그룹마다 별도의 B 트리를 구축하는 방식을 사용한다. 따라서 최장 빈번 패턴의 길이가 k인 경우, k 개의 B 트리가 구축된다. 또한, 질의 패턴과 대응되는 B 트리의 위치를 쉽게 파악하기 위하여 인덱스 테이블을 사용한다. 인덱스 테이블은 k개의 엔트리들을 가지며, 각각의 엔트리는 해당 길이와 대응되는 B 트리의 루트 노드의 위치를 저장한다. 인덱스 테이블은 그 크기가 일반적으로 매우 작으므로, 주기억장치 내에 상주가 가능하다.

그림 4는 최장 빈번 패턴의 길이가 k인 경우의 제안된 빈번 패턴 베이스와 원본 데이타베이스의 구축 상황을 예로 나타낸 것이다. 예를 들어, 빈번 패턴 CA는 길이가 2이므로 인덱스 테이블 내의 엔트리 F2가 가리키는 B 트리 내에 저장됨을 볼 수 있다. 또한, CA는 두 개의 주식 시퀀스 내에서 발생되며, 첫 번째 시퀀스 내의 세 곳에서, 두 번째 시퀀스 내의 두 곳에서 발생하였음을 볼 수 있다. 이러한 빈번 패턴 베이스의 구성 방식으로 인하여 관심 종목의 변화 패턴과 매치되는 빈번 패턴은 매우 신속하게 파악될 수 있다.

4. 규칙 매칭 및 추천

본 장에서는 빈번 패턴 베이스를 이용하여 투자자의 관심 종목의 주가 흐름과 매치되는 규칙을 찾아내고, 이를 기반으로 투자 유형을 추천하는 방법을 기술한다. 제 4.1절에서는 제안하는 방법의 기본 개념을 설명하고, 제 4.2절에서는 규칙을 매칭하고 투자 유형을 추천하는 구체적인 절차에 관하여 기술한다.

4.1 기본 개념

본 논문에서 제안하는 기법은 크게 두 가지 과정으로 구성된다. 첫 번째 과정은 규칙 매칭 과정으로서 주어진 관심 종목의 주가 변화 패턴과 매치되는 규칙 헤드를 빈번 패턴 베이스로부터 찾아낸다. 두 번째 과정은 투자 유형 추천 과정으로서 규칙 매칭 과정을 통하여 얻어진 규칙의 바디를 참조하여 해당 주식에 대한 투자 유형을 투자자에게 추천한다. 이때, 투자 유형은 매치된 규칙 헤드를 지지하는 기존의 주가 패턴의 이후의 변화 경향이 투자자의 투자 성향에 얼마나 부합되는가에 따라 결정된다.

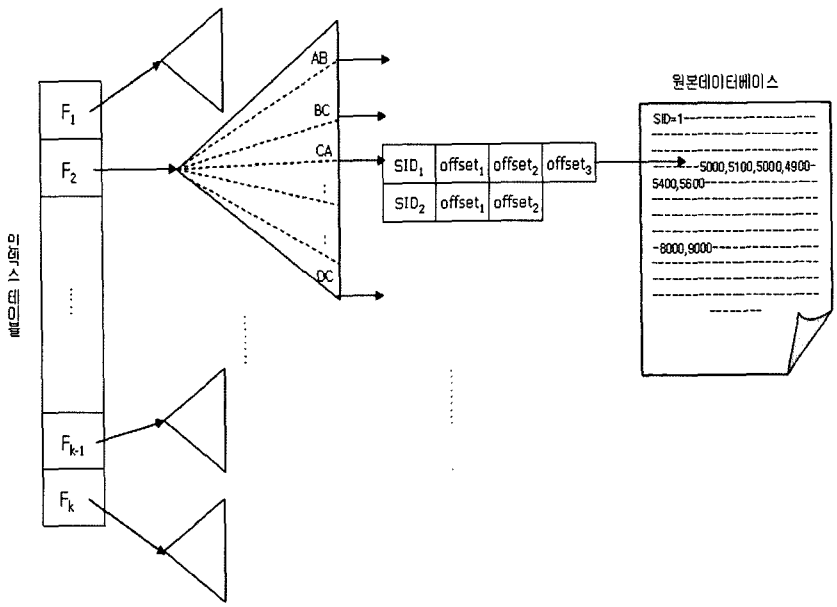

그림 4 빈번 패턴의 최대 길이가 k인 빈번 패턴에 대한 인덱스 구조

투자 유형 추천을 요청하는 질의는 관심 종목 및 투자 성향 등의 정보를 포함하며, 아래와 같은 형태로 제시된다.

Q = (item, QP, t, bodyLen, [minHold, maxHold], minConfidence)

여기서, item은 투자자의 관심 종목이며, QP(= <q[0], q[1], q[2], ..., q[Len(QP)-1]>)는 이 종목의 최근 변화 패턴을 시퀀스 형태로 표현한 것이다. t는 그림 2.1에 나타난 규칙 헤드의 끝 위치와 규칙 바디의 시작 위치 사이의 시간 간격을 의미한다. bodyLen은 규칙 바디의 길이이며, 추천을 위하여 QP와 매치되는 각 빈번 패턴의 시간 간격 t 이후의 길이 bodyLen 만큼의 주가 변화를 참조함을 의미한다. [minHold, maxHold]는 투자 유형을 결정하기 위한 상승률의 임계 범위를 의미한다. 투자 유형은 QP와 매치되는 규칙 헤드의 마지막 주가와 규칙 바디의 평균 주가 사이의 아래와 같이 정의되는 평균 상승률에 의하여 결정된다.

$$\text{평균 상승률}(H,B) = \frac{B\text{의 평균 주가} - H\text{의 마지막주가}}{H\text{의 마지막 주가}} \times 100$$

투자 유형은 이 평균 상승률 값에 따라 '매수', '매도', '보유', '무추천' 등을 고려할 수 있다. 즉, 평균 상승률이 maxHold 보다 크다면 '매수', minHold와 maxHold 사이라면 '보유', minHold 보다 작다면 '매도'의 투자 유형을 투자자에게 추천한다.

예를 들어, QP와 매치되는 규칙 헤드의 각 요소 값이 <5000, 5100, 4900, 5000>이고, 마지막 요소 값 5000의 발생한 이후 시간 간격 t 후에 나타나는 세 요소 값이 <5600, 5500, 5400>이라고 가정하자. 투자자 질의에 나타난 bodyLen이 3이라 할 때, 규칙 헤드의 마지막 주가는 5000이고, 규칙 바디 내의 평균 주가는 5500이 되므로 상승률은 10%가 된다. 이때, 투자자 질의에 나타난 [minHold, maxHold]가 [-5%, 5%]라 제시되었다면, 이 경우에는 해당 투자자에게 '매수'를 투자 유형으로서 추천하게 된다.

규칙이 가치를 가지기 위해서는 규칙 헤드를 지지하는 다수의 주가 변화 패턴들이 규칙 바디 구간 내에서 일정한 경향을 보여야 한다. minConfidence는 바로 이러한 일정한 경향을 보이는 수준을 정의하기 위한 최소 신뢰도를 나타낸다. 즉, 규칙의 신뢰도 confidence(H,B)가 minConfidence 보다 작으면, 규칙은 성립되지 않는 것으로 간주된다. 하나의 빈번 패턴이 서로 다른 두 가지 이상의 투자 유형을 동시에 추천하는 것을 방지하기 위하여 이 minConfidence가 50% 이상으로 설정되어야 한다.

여기서, 유의해야 할 점은 이와 같이 제시한 질의 모델은 하나의 예라는 것이다. 즉, 위의 질의 모델은 제안하는 기법의 동작 개념을 이해시키기 위한 것이며, 실제 응용 환경이나 투자자의 성향에 따라 다양한 형태의 설정이 가능하다. 예를 들어, 위의 질의 예제에서 나타난 t, bodyLen, minHold, maxHold, minConfidence 등의 값은 투자자의 성향에 따라 다르게 설정할 수 있다. 뿐만 아니라, 투자 유형을 결정하는 근거로서 평균 상승률이 아닌 다른 값이 사용될 수도 있다. 예를 들면, 어떤 투자자는 다음과 같은 최저 상승률을 근거로 투자 유형을 추천하도록 요구할 수 있다. 이 결과, 좀 더 보수적인 투자가 가능해진다.

$$\text{최저 상승률}(H,B) = \frac{B\text{의 최저 주가} - H\text{의 마지막주가}}{H\text{의 마지막 주가}} \times 100$$

즉, 본 논문에서 제시하는 기법은 응용 환경 및 투자자의 성향에 따라 다양한 설정을 가능하게 하는 투자 유형 추천 시스템을 위한 기본 프레임워크이다.

4.2 규칙 매칭 및 추천을 위한 처리 절차

본 절에서는 규칙 매칭 및 추천을 위한 처리 절차를 구체적으로 기술한다. 전체 과정은 아래와 같은 일곱 개의 단계로 구성된다.

단계 1. 질의 패턴의 심볼화

질의 패턴 QP(=<q[0], q[1], q[2], ..., q[Len(QP)-1]>)의 심볼화는 제 3.1.1절에서 언급한 주식 시퀀스의 사전 처리 방식을 그대로 이용한다. 먼저, 질의 패턴 QP를 증감 비율 시퀀스 QP′(= <q′[0], q′[1], q′[2], ..., q′[Len(QP)-2]>)로 표현한 후, 도메인 분류 방식에 의하여 이 증감 비율 시퀀스를 심볼 시퀀스로 변환한다. 이러한 심볼 시퀀스를 QP″(= <q″[0], q″[1], q″[2], ..., q″[Len(QP)-2]>)라 표기한다.

단계 2. 규칙 헤드 검색

변환된 심볼 시퀀스 QP″와 매치되는 규칙 헤드 RH를 빈번 패턴 베이스에서 검색한다. 이를 위하여 먼저 인덱스 테이블을 참조하여 QP″의 길이와 대응되는 B-트리의 루트 위치를 찾아낸다. 또한, 이 B-트리를 탐색하여 QP″와 매치되는 RH를 찾고, 이를 지지하는 각 주가 변화 패턴에 대하여 <SID, offset>의 쌍을 얻는다. 여기서, SID는 해당 주가 변화 패턴을 가지는 시퀀스 번호를 의미하며, offset은 이 시퀀스 내에서 해당 주가 변화 패턴의 시작 위치를 나타낸다.

단계 3. 규칙 헤드와 바디에 해당되는 실제 주가 검색

단계 2에서 얻은 <SID, offset>의 쌍을 이용하여 해당 규칙 헤드 RH와 매치되는 주가 변화 패턴들에 대하여 RH와 RB에 해당되는 시간 구간 내 실제 주가들을 원본 데이타베이스를 참조하여 구한다. 이 결과, RH와 RB에 해당되는 시간 구간 내 실제 주가들을 발견할 수 있다. 발견된 주가 변화 패턴의 수가 n일 때, 각 실제 주가 변화 패턴을 <RHi, RBi> (0<=i<n)이라 표기한다.

단계 4. 평균 상승률 계산

규칙 헤드와 매치되는 각 주가 변화 패턴에 대하여 투자 유형 추천의 근거로 사용되는 평균 상승률을 계산한다. 즉, <RHi, RBi> (0<=i<n)에 대하여 규칙 헤드의 마지막 주가 RH-endi와 규칙 바디의 RB-starti와 RB-endi사이의 평균 주가를 이용하여 평균 상승률을 계산한다.

단계 5. 신뢰도 계산

가치 있는 규칙인가의 여부를 판별하기 위하여 해당 규칙이 상승, 하락, 보합 중 어떠한 향후 경향을 지지하는지를 검증해야 한다. 이를 위하여 해당 규칙에 대한 신뢰도를 계산한다. 먼저, 규칙 헤드와 매치되는 각 주가 패턴이 발생한 이후에 나타나는 주가 변화를 투자자의 질의에 나타난 [minHold, maxHold] 값과 비교하여 상승, 하락, 보합인지의 여부를 판단한다. 즉, 해당 주가 변화 패턴의 평균 상승률이 minHold 보다 작다면, 이후의 주가가 하락한다고 간주한다. 유사한 방식으로 평균 상승률이 maxHold 보다 크다면 상승을, minHold와 maxHold 사이라면 보합을 지지하게 된다. 해당 규칙 헤드와 매치되는 모든 주가 변화 패턴들의 향후 경향이 파악되면, 상승, 하락, 보합에 대한 각각의 신뢰도를 각각 계산한다.

단계 6. 규칙 생성

단계 5에서 상승, 하락, 보합에 대하여 구한 신뢰도가 투자자 질의에 나타난 minConfidence 값 이상인 것을 규칙 바디로 하여 규칙을 생성한다. 이때, 상승, 하락, 보합에 대한 신뢰도가 모두 minConfidence 이하인 경우에는 규칙이 아닌 것으로 간주한다. 즉, 해당 빈번 패턴은 시간 구간 t 이후에 일정한 경향을 보이지 않는 것으로 간주하는 것이다.

단계 7. 투자 유형 추천

질의의 결과로 탐사된 규칙을 참조하여 투자 유형을 투자자에게 추천한다. 즉, 규칙이 주가가 상승 된다는 것을 지지하는 경우 '매수'를 추천하며, 하락 된다는 것을 지지하는 경우 매도를 추천한다. 또한, 보합이라는 것을 지지하는 경우, 투자자가 해당 종목을 가지고 있다면 '보유'를 추천한다. 단, 규칙이 탐사되지 않은 경우에는 투자 유형을 추천하지 않는다. 투자 유형 추천 시, 규칙의 지지도와 신뢰도를 함께 제시하여 투자자에게 규칙에 대한 신뢰성과 유용성을 제공한다.

그림 5에 질의 처리 과정의 예를 보인다. 우선 그림 5(a)에서 질의 시퀀스 QP=<5000, 5100, 4900, 5400>를 증감 비율로 표현 한 후, 이를 다시 도메인 분류 방식에 의해 심볼로 변환한 심볼 시퀀스 QP'=<DBF>를 생성한다. 다음, 그림 5(b)에서 변환된 심볼 시퀀스 QP'과 매치되는 규칙 헤드를 빈번 패턴 베이스에서 검색하고,

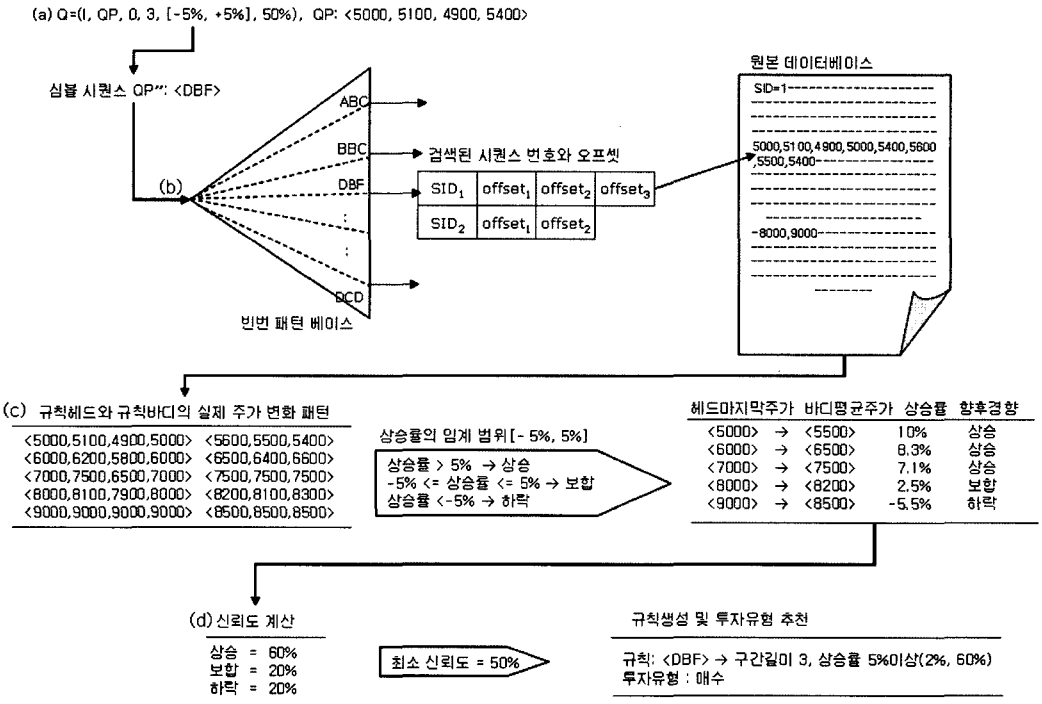

그림 5 질의 처리 과정의 예

이를 지지하는 각 주가 변화 패턴에 해당하는 <SID, offset>을 얻는다. 다음, 얻어진 <SID, offset>쌍을 이용하여 원본 데이타베이스에서 규칙 헤드와 규칙 바디에 해당되는 시간 구간 내 실제 주가들을 가져온다. 그림 5(c)에서는 규칙 헤드와 매치되는 5개의 <RH_i, RB_i>(0<=i<5)의 쌍으로 구성된 실제의 주가 변화 패턴을 발견하였으며, 발견된 각 주가 변화 패턴에 대하여 규칙 헤드의 마지막 주가와 규칙 바디 구간내의 평균 주가를 이용하여 평균 상승률을 계산하고, 각 주가 변화 패턴이 향후 경향을 파악한다.

그림 5(d)에서는 규칙의 신뢰도를 구하기 위하여 상승, 하락, 보합에 대한 각각 신뢰도를 계산한 후, 최소 신뢰도를 만족하는 것을 규칙 바디로 하여 규칙을 생성한다. 그림에서는 상승에 대한 신뢰도가 최소 신뢰도 50% 이상을 만족하므로, 규칙 헤드 'DBF'와 상승을 지지하는 조건(규칙 바디 길이 3, 상승률 5%이상)을 규칙 바디로 하는 규칙 'DBF->구간 길이 3, 상승률 5%이상'을 생성한다. 질의의 최종 결과로는 규칙을 참조하여 규칙이 주가가 상승한다는 것을 지지하므로 투자 유형으로 '매수'를 투자자에게 추천한다.

5. 성능 평가

본 장에서는 실험에 의한 성능 분석을 통하여 제안하는 기법의 우수성을 규명한다. 먼저 실험 환경을 설명한 후, 실험 결과를 분석한다.

본 연구에서는 실제의 주식 데이타를 이용한 실험을 통하여 성능을 분석한다. 주식 데이타로서 한국의 코스피(KOSPI) 주식 중 거래량이 많은 10개의 우량 종목을 선택하였으며, 각 종목에 대하여 1985년 5월 28일부터 2005년 5월 27일까지 20년간 일별 종가를 수집한 데이타를 사용한다[19]. 본 실험에서는 이 데이타를 KOSPI-Data라 부른다. KOSPI-Data의 종목별 주식 데이타는 5,557개의 실수 값들을 갖는 시퀀스로 이루어지며, 제 3.1.1절에서 보인 도메인 분류 방식에 의하여 심볼 시퀀스로 변환된 후 실험에 적용된다. 따라서 한 개의 심볼 시퀀스는 하루 동안의 주가 변화를 가리킨다.

실험을 위한 하드웨어 플랫폼으로는 512MB의 주기억장치와 7200RPM의 HDD를 장착한 Intel Pentium-IV 2.6GHz의 PC를 사용한다. 또한, 소프트웨어 플랫폼으로는 Redhat Linux 커널 버전 2.4를 운영 체제로 사용한다.

본 논문에서 제안하는 주식 데이타 예측 방법은 시계열 분석, 신경망 등을 이용한 기존의 주식 데이타 예측 방법과는 차이가 있다. 즉, 기존 방법과는 달리 제안하는 방법은 사용자의 투자 스타일을 충실하게 반영한 투자 추천을 제공할 수 있다는 중요한 특징을 가진다. 따라서 다른 방법과 성능을 직접 비교하는 것은 큰 의미가 없다. 본 성능 평가에서는 제안된 방식이 얼마만큼의 정확도를 보이며, 얼마나 빠르게 처리하는가를 보이는 것을 주 목표로 한다.

우선, 실험 1에서는 본 논문에서 제안한 규칙 탐사 및 매칭 모델의 검색 성능을 측정한다. 성능 평가 지수로서 예측 결과의 만족율과 추천율을 사용한다. 예측 결과의 만족율(satisfaction ratio)은 사용자가 질의 처리 결과로 얻어진 규칙의 추천 값에 대하여 실제로 어느 정도 만족하는지를 나타내는 비율이며, 아래와 같이 정의된다.

$$\text{만족율}(Q) = \frac{\text{투자자가 의 } Q\text{의 추천 값에 의한 투자 결과에 만족하는 횟수}}{\text{질의 } Q\text{에 대하여 추천 값이 응답된 횟수}}$$

따라서, 만족율이 높을수록 이 모델에서 추천한 추천 값이 실제로 많이 적중했다는 의미가 되며, 만족율을 통하여 제안한 방법의 신뢰성을 알 수 있다.

규칙 바디의 평균 상승률은 주식 투자자가 투자를 결정하는 매우 중요한 요인으로 작용한다. 주식 투자자는 보다 확실한 수익을 보장하는 종목을 투자 대상으로 선택하기 위하여, 추천의 기준으로 사용되는 평균 상승폭을 크게 지정한다. 그러나 투자자는 실제 자신이 투자한 종목의 상승폭이 이보다 작더라도 만족하는 경우가 대부분이다. 즉, 투자 대상의 선택을 위한 평균 상승폭과 투자자가 투자 후의 만족할 수 있는 상승폭은 차이가 있다.

본 실험에서는 이러한 상황을 반영하기 위하여 만족수준(satisfaction level)이라는 개념을 사용한다. 만족수준 sLevel은 규칙 생성 단계에서 사용한 평균 상승률의 임계 범위 [minHold, maxHold]를 기준으로 투자 결과에 의한 수익이 어떤 범위 내에 들어가야 투자자가 만족하는지를 결정하는 값을 의미한다. 본 실험에서는 투자 추천에 대한 만족율을 다음과 같은 방식에 의하여 계산한다. 우선, 규칙 생성 단계에서는 평균 상승률의 임계 범위 [minHold, maxHold]를 이용하여 빈번 패턴 베이스를 검색하여 규칙을 생성하고 투자 유형을 얻는다. 실험 데이타 검증 단계에서는 [minHold*sLevel, maxHold*sLevel]를 임계 범위로 사용하여 추천된 투자 유형에 대하여 만족하는지의 여부를 판단한다. 즉, 투자 후의 실제 상승폭이 투자 대상의 선택을 위한 평균 상승폭의 (sLevel*100) % 이상이 되면 투자자는 만족한다고 판단하는 것이다.

추천율(recommendation ratio)은 전체 사용자 질의의 몇 퍼센트가 규칙 탐사에 성공하여 투자 유형에 대한 추천을 얻게 되는지 나타내는 비율이며, 아래와 같이 정의된다.

$$\text{추천율}(Q) = \frac{\text{질의 } Q\text{에 대하여 추천 값이 응답된 횟수}}{\text{질의 } Q\text{가 발생된 횟수}}$$

예측 결과의 만족율 및 추천율은 규칙 탐사 및 매칭

모델을 구성하는 다양한 종류의 요소 값의 변화에 영향을 받는다. 이들은 종목별 주식 시퀀스를 심볼 시퀀스로 변형하는 과정에 정의되는 심볼의 개수, 빈번 패턴 생성에 사용된 최소 지지도, 규칙 탐사 과정에 적용되는 규칙 바디의 시작 위치, 규칙 바디의 길이, 규칙 바디의 평균 상승률의 임계 범위, 규칙 생성을 위한 최소 신뢰도 등이다.

빈번 패턴 베이스는 전체 KOSPI-Data의 70%에 해당하는 부분을 사용하여 구축하였으며, KOSPI-Data의 나머지 30%에 해당하는 부분으로부터 추출한 17,040개의 빈번 패턴을 질의 시퀀스로 사용하였다. 빈번 패턴의 기준을 위한 최소 지지도는 2%로 설정하였다. 또한, 투자 유형을 위한 추천 값으로는 매수와 매도만을 사용하였다. 또한, 이후의 실험에서 기본적으로 사용하는 각 요소의 값은 다음과 같다. 심볼의 개수는 3, 최소 지지도는 2%, 평균 상승률의 임계 범위는 ±2%, 규칙 헤드와 규칙 바디 사이의 시간 간격은 0, 규칙 바디의 길이는 1, 최소 신뢰도는 60%, 만족 수준은 60%이다.

그림 6은 규칙 바디의 평균 상승률의 임계 범위 증가에 대한 만족율과 추천율의 변화를 나타낸다. 실험 결과에 의하면 평균 상승률의 임계 범위가 ±1%에서 ±4%까지 증가함에 따라 만족율은 서서히 증가하다가 감소하는 것으로 나타났다. 또한, 추천 값에 대한 만족율은 평균적으로 약 80%의 높은 값을 보였다. 그러나 평균 상승률의 임계 범위가 증가함에 따라 추천율은 급격히 감소하는 것으로 나타났다. 특히, 평균 상승률의 임계 범위가 ±3%인 경우, 추천율은 3.3%에 그치는 것으로 나타났다.

앞서 기술한 바와 같이 추천율은 사용자가 추천을 받게 되는 비율을 의미한다. 추천율이 아주 높을 필요는 없으나, 0에 가까운 경우는 추자 대상을 거의 추천하지 않는다는 것을 의미하므로 실질적으로 도움이 되지 않는다.

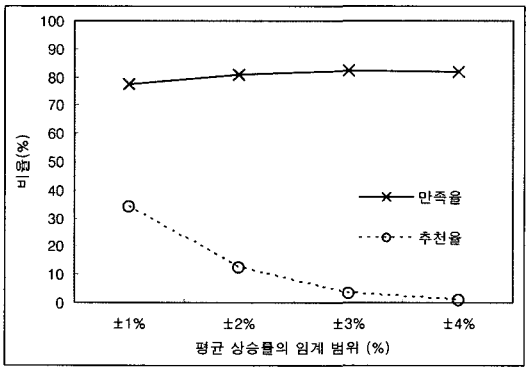

그림 6 평균 상승률의 임계 범위 변화에 따른 만족율과 추천율의 변화

본 연구에서는 적절한 추천율과 높은 만족율을 갖는 주식 투자 시스템 개발을 목적으로 한다. 실험 결과에 의하면 ±2%의 평균 상승률을 임계 범위로 사용하는 사용자 질의는 시스템으로부터 약 13%의 비율로 매도 혹은 매수의 추천 값을 제공 받게 되며, 이 추천 값에 대하여 약 80%의 만족율을 기대할 수 있음을 알 수 있다.

그림 7은 도메인 분류에 사용되는 심볼 개수의 변화에 따른 만족율과 추천율의 변화를 나타낸다. 심볼의 개수가 3, 5, 7, 9인 경우, 만족율은 76~83% 정도로 나타났으며, 심볼의 개수에 큰 영향을 받지 않는 것으로 나타났다. 한편, 추천율은 심볼의 개수가 3인 경우에 약 13%를 나타내며, 심볼의 개수가 증가함에 따라 그 값이 근소하게 감소하는 것으로 나타났다. 실험 결과에서 나타난 최적의 심볼 개수는 3이며, 약 80%의 만족율과 약 13%의 추천율을 보였다.

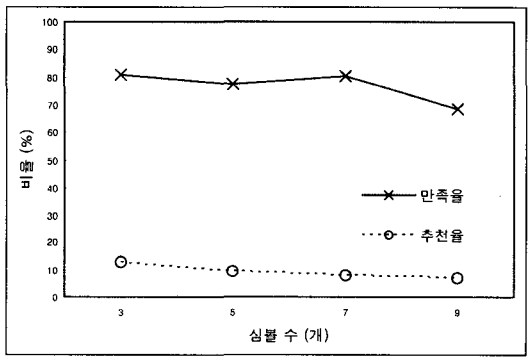

그림 7 심볼 개수의 변화에 따른 만족율과 추천율의 변화

그림 8은 규칙 헤드와 규칙 바디 사이의 시간 간격 t의 변화에 따른 만족율과 추천율의 변화를 나타낸 것이다. 시간 간격이 증가할수록 만족율과 추천율이 감소하는 것을 볼 수 있다. 이와 같은 결과는 규칙 헤드와 규칙 바디 사이의 시간 간격이 증가할수록 둘 간의 연관성이 적다는 점으로부터 쉽게 유추할 수 있다. 실험 결과로부터 시간 간격 t가 0인 경우, 최적의 만족율과 추천율을 얻을 수 있다는 것을 알 수 있다.

그림 9는 규칙 바디의 길이 변화에 따른 만족율과 추천율의 변화를 나타낸 것이다. 실험 결과로부터 규칙 바디의 길이가 길어짐에 따라 만족율과 추천율이 서서히 감소하고 있음을 볼 수 있다. 이와 같은 결과는 규칙 바디의 길이가 증가할수록 규칙 바디의 뒷부분의 값들은 규칙 헤드와의 연관성이 적어진다는 사실로부터 쉽게 해석될 수 있다. 실험 결과에 따르면, 규칙 바디의 길이가 1일인 경우 최적의 만족율과 추천율을 얻게 되는 것으로 나타났다.

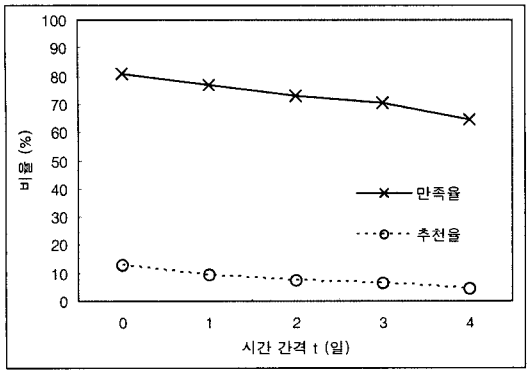

그림 8 규칙 헤드와 규칙 바디 사이의 시간 간격 t의 변화에 따른 만족율과 추천율의 변화

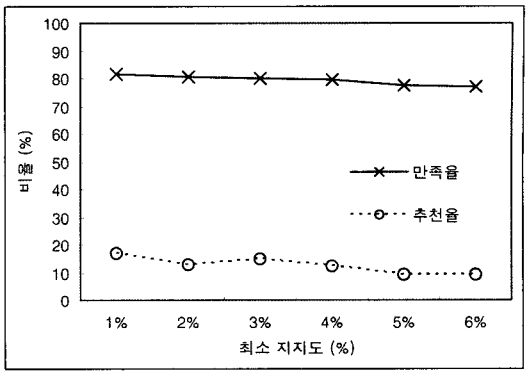

그림 10 빈번 패턴 생성에 사용된 최소 지지도의 변화에 따른 만족율과 추천율의 변화

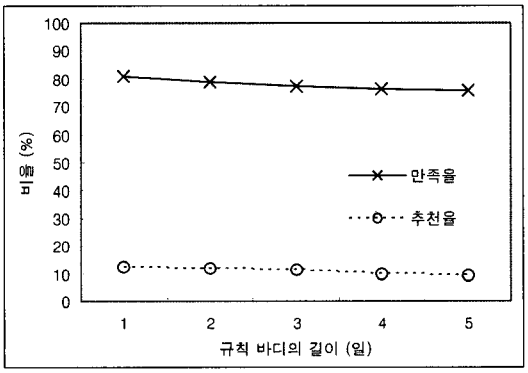

그림 9 규칙 바디의 길이 변화에 따른 만족율과 추천율의 변화

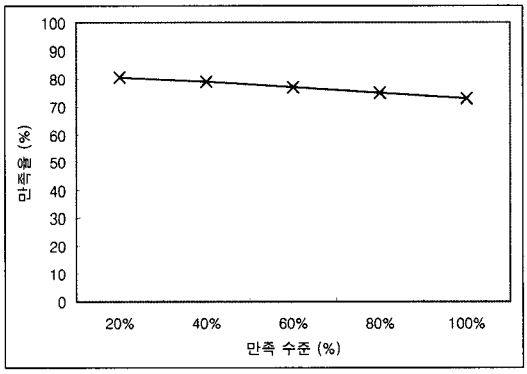

그림 11 만족율 검증에 사용되는 만족 수준의 변화에 따른 만족율 변화

그림 10은 빈번 패턴 생성에 사용된 최소 지지도의 변화에 따른 만족율과 추천율의 변화를 나타낸 것이다. 실험 결과로부터 최소 지지도가 증가함에 따라 만족율은 서서히 감소하고 있음을 볼 수 있다. 한편, 추천율은 최소 지지도가 증가함에 따라 부분적인 증감의 변화를 보이지만 역시 서서히 감소하고 있음을 알 수 있다. 빈번 패턴 생성을 위한 최소 지지도가 증가하게 되면 저장하는 빈번 패턴 베이스의 크기가 줄어들며, 이것이 만족율과 추천율의 감소로 이어지는 것이다.

끝으로, 그림 11은 만족율을 검증할 때 사용한 만족 수준의 변화에 따른 만족율의 변화를 나타낸 것이다. 만족 수준이 1.0인 경우에는 투자 시점에 주어진 상승률 임계 범위 자체를 만족해야 규칙이 적중한 것으로 간주한다. 반면, 만족 수준이 감소할수록 검증에서 사용되는 보유 변동률의 범위가 감소하게 된다. 실험 결과, 만족 수준이 증가할수록 만족율이 감소하고 있으나, 만족 수준이 100%일 경우와 비교했을 때 만족율의 차이는 최대 7%정도로 크지 않은 것으로 나타났다. 또한, 모든 경우에서 70%이상의 높은 만족율을 나타냄으로써 제안된 기법에 의한 추천에 대한 만족의 정도가 큰 것으로 나타났다.

이와 같은 만족율과 추천율이 통계적으로 얼마나 의미가 있는 것인지 알아보자. 추천율은 어떤 빈번 발생 패턴이 현재 주가 시퀀스에서 발견되었을 경우, 제안한 방법에 의하여 'BUY'나 'SELL'을 추천할 확률이므로, 제안한 방법의 정확도와는 직접적인 관계가 없다. 한편 만족율은 실제 주가 시퀀스를 대상으로 제안한 방법을 실행했을 때 적중한 비율을 가리키며, 따라서 제안한 방법이 얼마나 효과적인지를 실험으로 검증할 수 있는 중요한 지표이다. 이를 확인하기 위하여, 빈번 발생 패턴이 현재 주가 시퀀스에서 발견되었을 때, 추천값을 랜덤하게 질의를 처리하는 실험을 수행하고 그 결과로 나온 만족율을 제안한 방법의 만족율과 비교하였다. 그 결과, 우연에 의해 제안한 방법의 정확도를 얻을 확률(p-value)은 0.1% 이하로 계산되었으며, 따라서 제안한 방법이 실제로 효과적임을 확인할 수 있다.

실험 2에서는 본 논문에서 제안한 기법의 성능 평가를 위하여 빈번 패턴 베이스 생성 시간과 질의 처리 시

간을 분석한다. 빈번 패턴 베이스를 구축하기 위하여 전술한 KOSPI-Data를 사용하였다. 질의 시퀀스는 최소 지지도 2%의 모든 빈번 패턴을 사용하였으며, 질의 처리 시 매수와 매도만을 추천 값으로 인정하였다. 또한, 심볼의 개수는 3, 최소 지지도는 2%, 평균 상승률의 임계 범위는 ±2%, 규칙 헤드와 규칙 바디 사이의 시간 간격은 0, 규칙 바디의 길이는 1, 최소 신뢰도는 60%로 설정하였다.

그림 12는 최소 지지도 변화에 따른 빈번 패턴 베이스 생성 시간의 변화를 나타낸 것이다. 최소 지지도가 증가함에 따라 빈번 패턴 베이스 구축 시간이 점차 감소하고 있음을 알 수 있다. 이것은 최소 지지도가 증가할수록 데이타 시퀀스로부터 추출되는 빈번 패턴의 수는 적어지기 때문이다.

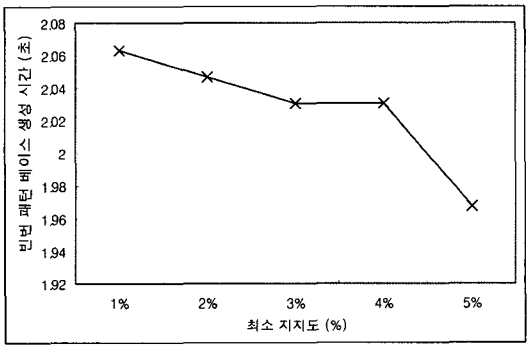

그림 12 최소 지지도 변화에 따른 빈번 패턴 베이스 생성 시간의 변화

그림 13은 질의 시퀀스의 길이 변화에 따른 질의 처리 시간의 변화를 나타낸 것이다. 질의 시퀀스의 길이가 증가함에 따라 질의 처리 시간이 감소하는 것을 볼 수 있다. 그 이유는 빈번 패턴 베이스를 구성하는 빈번 패턴의 분포에 있어 빈번 패턴의 길이가 증가할수록 그 출현 횟수가 감소하기 때문이다. 즉, 길이가 긴 질의에 의하여 탐색되는 빈번 패턴은 길이가 짧은 질의에 의하여 탐색되는 빈번 패턴보다 발생 빈도가 낮으며, 따라서 질의 처리 시간은 질의 길이가 길어짐에 따라 감소하는 것이다.

그림 14는 데이타 크기 변화에 따른 빈번 패턴 베이스 생성 시간과 질의 처리 시간의 변화를 나타낸다. 크기가 다른 데이타 생성을 위하여 KOSPI-Data를 2배, 3배, 4배로 복사하여 사용하였다. 가로축의 1S는 KOSPI-Data를 나타내며, 2S, 3S, 4S는 각각 KOSPI-Data를 2배, 3배, 4배로 복사한 데이타를 나타낸다. 이와 같이 복사에 의하여 데이타 크기를 변화시키는 방식을 사용하는 이유는, 동일한 최소 지지도와 최소 신뢰도를 이용한

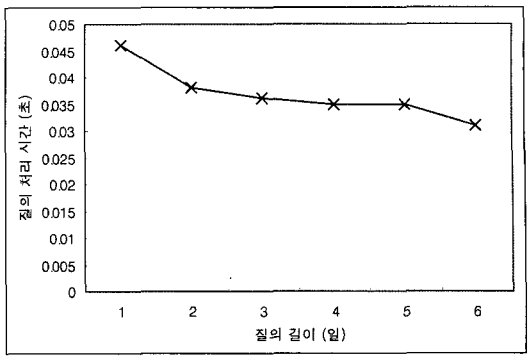

그림 13 질의 길이 변화에 따른 질의 처리 시간의 변화

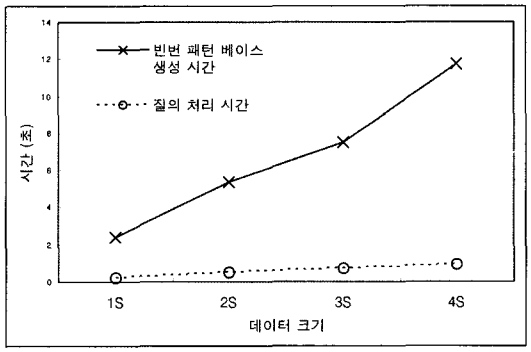

그림 14 데이타 크기 변화에 따른 빈번 패턴 베이스 생성 시간과 질의 처리 시간

질의를 다른 크기의 데이타에 적용시켰을 때 질의 결과의 만족율이 변하지 않도록 하기 위한 것이다. 실험 결과로부터 데이타 크기가 증가함에 따라 빈번 패턴 베이스 생성 시간과 질의 처리 시간이 거의 비례하여 증가함을 볼 수 있다. 이는 동일한 데이타가 2배로 그 크기가 증가하면 빈번 패턴의 발생 회수도 똑같이 2배로 늘어나며, 따라서 빈번 패턴 베이스 생성 시간과 질의 처리 시간도 2배로 증가하기 때문이다.

6. 결 론

본 논문에서는 주식 데이타베이스로부터 규칙을 탐사함으로써 투자자에게 주식 투자 유형을 추천해 주는 방안에 관하여 논의하였다. 제안된 방안에서는 탐사된 규칙들을 대상으로 하는 투자자의 질의를 통하여 향후 이 주식의 가격 추이를 예측하고, 이를 기반으로 투자자에게 주식의 보유, 매수, 매도 등의 투자 행위를 추천한다.

이를 위하여 본 논문에서는 먼저 주식 투자 유형의 추천을 위한 새로운 규칙 모델을 정의하였다. 제안된 규칙 모델에서는 빈번하게 발생하는 주가 변화 패턴의 이후의 주가 변화 경향이 투자자의 투자 조건과 매치하는

경우 이를 규칙으로 생성한다. 본 연구에서는 규칙 헤드는 투자자의 특성에 별다른 영향을 받지 않는 반면, 규칙 바디에 대한 조건은 투자자마다 다는 점에 착안하여 규칙 탐사 과정에서 전체 규칙이 아닌 규칙 헤드들만을 탐사하여 저장해 두는 새로운 방식을 제안한다. 이 결과, 투자자 별로 달라질 수 있는 규칙 바디에 대한 조건을 유연하게 정의하는 것을 허용하며, 규칙의 수를 줄임으로써 전체 규칙 탐사 성능을 개선할 수 있다. 효율적인 규칙 탐사와 매칭을 위하여 빈번 패턴들을 효과적으로 탐사하는 방법, 빈번 패턴 베이스를 구축하는 방법, 그리고 이들을 인덱싱하는 방법을 제안하였다. 또한, 투자자의 질의가 발생하는 경우, 빈번 패턴 베이스로부터 이와 매치되는 규칙을 발견하고, 이 결과를 이용하여 투자자에게 투자 유형을 추천해 주는 방법을 제안하였다.

제안된 기법의 우수성을 규명하기 위하여 추천 값에 대한 만족율과 응답 시간에 대한 다양한 실험을 수행하였다. 실험 결과에 의하면 제안된 기법은 대부분의 경우 70~80% 내외의 추천 값에 대한 만족율과 온라인 환경의 사용자가 기다릴 수 있는 1초 이내의 응답 시간을 보였다.

향후에는, 제안된 모델을 다수의 사용자가 동시에 사용하는 환경에서 수많은 질의들을 실시간으로 빠르게 처리하는 방법과, 최근에 발견된 패턴이 과거의 패턴보다 더 큰 가중치를 갖도록 하여 최근의 주가 변화 동향을 더 잘 반영하도록 하는 방법에 대한 연구를 수행할 예정이다.

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 ICDE, 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, Jul. 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 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] H. White, "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns," In Proc. IEEE Int'l. Conf. on Neural Networks, pp. II451-II458, 1988.

[10] E. Saad, D. Prokhorov, and D. Wunsch II, "Comparative Study of Stock Trend Prediction Using Time Delay, Recurrent and Probabilistic Neural Networks," IEEE Trans. on Neural Networks, pp. 1456-1470, 1998.

[11] B. Wah and M. Qian, "Constrained Formulations and Algorithms for Stock-Price Predictions Using Recurrent FIR Neural Networks," AAAI/IAAI pp. 211-216, 2002.

[12] G. Das, K.-I. Lin, H. Mannila, Gopal Renganathan, and Padhraic Smyth, "Rule Discovery from Time Series," In Proc. Int'l. Conf. on Knowledge Discovery and Datamining, pp. 16-22, 1998.

[13] S. Park and W. W. Chu, "Discovering and Matching Elastic Rules From Sequence Databases", in Fundamenta Informaticae, Vol, 47, No. 1-2, pp. 75-90, Aug-Sept, 2001.

[14] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," Information Systems Vol. 26, No. 1, pp. 35-58, 2001.

[15] W. W. Chu and K. Chiang, "Abstraction of High Level Concepts from Numerical Values in Databases," In Proc. AAAI Workshop on Knowledge Discovery in Databases, pp. 133-144, 1994.

[16] J. Han, M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2001.

[17] R. Agrawal, R. Srikant, "Fast Algorithms for mining Association Rules," In Proc. Int'l. Conf. on VLDB, pp. 487-499, 1994.

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

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

하 유 민

1997년 3월~2005년 2월 연세대학교 컴퓨터과학과(학사). 2005년 3월~2007년 2월. 연세대학교 컴퓨터과학과(석사). 관심분야는 데이타 마이닝, 스트림 데이타베이스, 멀티미디어 데이타베이스 등

김 상 욱

정보과학회논문지 : 데이타베이스
제 34 권 제 2 호 참조

원 정 임

정보과학회논문지 : 데이타베이스
제 34 권 제 2 호 참조

박 상 현

정보과학회논문지 : 데이타베이스
제 34 권 제 2 호 참조

윤 지 희

정보과학회논문지 : 데이타베이스
제 34 권 제 2 호 참조