FP-tree를 이용한 유사 패턴 주식종목 추출

심종보*, 김원영*, 김응모*
*성균관대학교 컴퓨터공학과
e-mail:hantough11@hotmail.com

Mining of Stocks Having Similar Pattern using FP-Tree

Jong-Bo Sim*, Won-Young Kim*, Ung-Mo kim*
*School of Information and Communication Engineering, Sungkyunkwan University

요 약

최근 컴퓨터와 인터넷의 발달로 과거 창구거래를 이용하던 방법에서 HTS(Home Trading System)을 이용하여 거래하게 됨으로써 개인투자자들도 쉽게 주식투자를 할 수 있게 되었다. 그러나 개인들이 방대한 양의 과거 데이터를 분석하기에는 상당한 어려움이 있다. 본 논문에서는 주식 데이터베이스로부터 과거 특정 종목들 간 연관성을 추출하여 투자자들로 하여금 주식 선별에 참고가 될 수 있는 방안에 관하여 논의한다. 기존의 논문에서 제안된 과거 패턴을 이용하여 미래의 주가변화를 예측하는 것과 달리, 종목들 간에 연관성을 통하여 하나의 테마가 형성 되었을 때 주도주의 변화로 관련주의 변화를 파악하여 투자에 유익한 정보를 제공하는데 목적이 있다.

1. 서론

최근 국내의 증권거래는 거래를 하기 위해 증권사로 직접 찾아가야 했던 과거와는 달리, 가정용 투자정보시스템에서 발전된 HTS(Home Trading System)가 널리 보급되면서 빠른 속도로 성장하고 있다[1]. 또한 HTS의 발달로 개인투자자들의 직접투자가 증가하고 있다. 하지만 개인들이 방대한 양의 과거 데이터를 분석하여 유용한 정보를 얻기에는 상당한 어려움이 따른다.

이미 주가의 변화를 예측하는 다양한 방법이 존재하지만, 본 논문에서는 특정 종목과 유사한 패턴을 갖는 종목들을 발굴하고, 특정 종목의 주가 변화와 비교하여 유사 종목들의 주가를 예측하는 방법을 제안한다. 이를 위해, 본 논문에서는 FP-tree라는 데이터 마이닝에서 빈발 패턴을 찾는데 사용되는 방법을 이용해 유사한 패턴을 갖는 종목 추출 방법을 제안한다[2].

본 논문은 다음과 같이 구성된다. 2장에서는 본 논문과 관련된 데이터마이닝 기법의 기본적인 내용을 설명하고, 3장에서는 데이터마이닝 기법을 적용하여 유사한 패턴의 종목을 추출하는 기법에 대해 설명한다. 4장에서는 결론과 함께 향후 과제를 제시하고 끝을 맺는다.

2. 관련 연구

2.1 데이터 마이닝(Data Mining)

데이터 마이닝은 대량의 데이터로부터 새로운 지식을 추출하는 것이다[3]. 즉 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아내는 것이다.

2.2 연관 규칙(Association Rule)

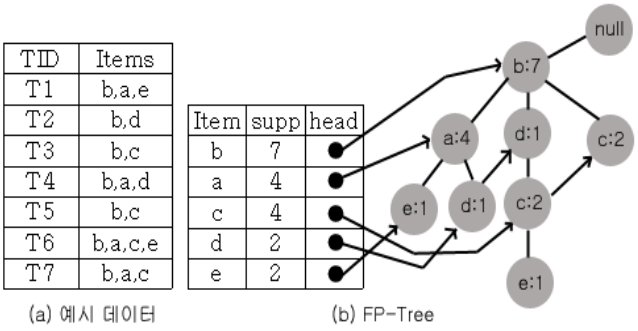
연관규칙은 대규모 데이터 항목의 집합 사이에서 유용한 연관성과 상관관계를 찾는 방법이다[4]. 간단히 말하면 데이터의 항목들 간의 조건-결과 식으로 표현되는 유용한 패턴을 말한다. 예를 들면 슈퍼마켓에서 X를 구매하는 고객은 Y도 동시에 구매한다는 연관성을 알아냈다고 한다. 이 때 조건은 'X'이며, 결과는 'Y'라 할 수 있다. 여기서 X, Y와 같이 고객이 구매하는 물품들의 집합을 트랜잭션이라 하고 X→Y의 형태로 표현한다. 연관규칙에 대한 객관적 척도로는 주어진 규칙이 만족 시키는 트랜잭션들의 퍼센트를 나타내는 규칙 지지도와 탐색된 연관성의 확실성 정도를 평가해 주는 신뢰도가 있는데 다음과 같이 정의 된다.

$$\text{지지도}(X \rightarrow Y) = P(X \cup Y)$$

$$\text{신뢰도}(X \rightarrow Y) = P(X \mid Y)$$

2.3 FP-Tree

FP-tree 연관 규칙 탐사 알고리즘은 DB 스캔에 대한 부담을 획기적으로 절감시킴으로써 전체적인 성능을 향상시키고자 제안되었고, 따라서 다른 기법에 기반 하는 알고리즘 보다 성능이 매우 우수한 것으로 알려져 있다[5]. FP-tree 알고리즘은 분할 정복 기법을 사용하고, 빈발 항목을 가지는 데이터베이스를 FP-tree로 압축하여 사용한다. FP-tree 알고리즘은 크게 생성과 탐색 단계로 이루어져 있다. (그림 1)은 예시 데이터와 FP-tree이다.



(그림 1) FP-tree 알고리즘 예시 데이터

3. 유사 패턴 주식종목 추출

3.1 제안 배경

주가는 개별적으로 변화하기도 하지만 서로 연관성이 있는 종목들은 대부분 비슷한 패턴을 가지고 움직인다. 예를 들면 LED관련주, 대체에너지관련주 등과 같이 동일한 테마를 갖고 있거나, 같은 계열사의 종목일 경우, 또는 같은 업종일 경우가 그렇다. 따라서 이들을 대표하는 주도주가 있는 반면, 주도주를 따라 가는 관련주가 존재하기 마련이다. 즉 주도주의 추세를 파악하고, 연관성이 있는 관련주들을 추출하여 예측 할 수 있다면 투자의 폭이 좀 더 확대될 것이라 예상된다.

3.2 제안 사항

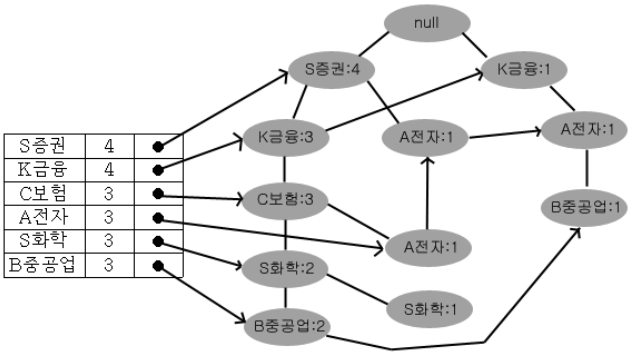
먼저 기준이 되는 종목 A를 선정하고, 탐사 횟수를 줄이기 위해 동일한 테마, 업종 등과 같이 최소한의 연관성이 있는 종목들을 선별한다. 이는 각 증권회사의 HTS에서 기본적으로 제공되고 있는 기능이므로 쉽게 선별이 가능하다. 선별된 데이터를 가지고 FP-tree 알고리즘을 이용하여 유사한 패턴을 가지는 종목을 추출한다. 이 때, 종목 A의 변동 폭에 임의의 오차범위 내에서 움직이는 종목을 항목집합에 포함시킴으로써 좀 더 정확성을 높일 수 있다.

TID	A의 등락	오차 범위 2% 이내의 항목집합
090102	+10%	S증권, K금융, C보험, A전자, S화학, B중공업
090103	-8%	S증권, K금융, C보험, A전자, S화학
090104	+4%	S증권, A전자, L제과
090105	+2%	K금융, C보험, B중공업, L제과
090106	-3%	S증권, K금융, C보험, S화학, B중공업, H건설

<표 1> 예제 데이터[종가 등락폭 기준]

<표 1>은 1주일간의 종목 A의 주가 변동과 임의의 오차범위 2% 이내로 등락폭이 같은 종목들의 데이터이다. 여기서 TID는 거래날짜를 의미한다. 최소지지도를 3이라 하고, 이 데이터를 FP-tree알고리즘에 적용한다.

(그림 2)는 <표 1>에 대한 결과로 만든 FP-tree이다. 위의 자료를 이용하여 투자자는 A종목이 상승 추세일 때에 한에 S증권부터 B중공업까지 투자할 수 있는 기회를 가질 수 있게 된다.



(그림 2) 표 1에 대한 FP-tree

4. 결론

본 논문에서는 데이터마이닝 기법을 이용하여 유사 패턴을 갖는 종목을 추출하는 방법을 제안하였다. 이 방법을 통하여 투자자는 좀 더 다양한 방법으로 투자에 접근 할 수 있을 것으로 생각된다. 그러나 이 방법은 기존의 다른 주식 투자 방법들과 마찬가지로 완벽한 투자방법은 될 수 없다. 하지만 다른 방법과 함께 사용한다면 좀 더 효율적으로 투자를 할 수 있을 것이라 기대한다. 이와 관련하여 향후 연구 과제로는 등락폭 뿐 아니라 다른 기술적 지표를 포함시킴으로서 좀 더 정확성을 높이는 연구가 필요하다.

감사의 글

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-007 5771).

참고문헌

[1] 이윤영 외 7명 “홈트레이딩 시스템 서비스의 보안 취약점 분석 및 평가기준 제안” 2008. 2. 정보보호학회론문지 제 18권 제 1호.

[2] 박용기, 김황수 “FP-Tree를 이용한 문서 분류”, 2005. 한국컴퓨터종합학술대회 논문집 제 32권 제 1호.

[3] 이기욱, 성창규 “데이터 마이닝 기법을 이용한 추천 시스템의 구현” 2006. 3. 한국컴퓨터정보학회 논문지 제 11권 제 1호.

[4] 박우창, 승현우, 용환승, 최기헌, “데이터 마이닝 개념 및 기법”, 자유아카데미, 2000

[5] 이형봉, 김진호, “FP-tree와 DHP 연관 규칙 탐사 알고리즘의 실험적 성능 비교”, 2008. 6. 정보과학회논문지 제 35권 제 3호.