MST 알고리즘을 이용한 대용량 주식 정보 분석 모듈

허 준, 권오규
국가수리과학연구소 수리모델 연구부
e-mail:heojoon@nims.re.kr

An Analysis Module of Mass Stock Information using MST Algorithm

Joon Heo, Okyu Kwon
Division of Mathematical Models, National Institute for Mathematical Sciences

요 약

주식 데이터의 분석을 위해서는 전문적인 분석 알고리즘 지식, 주식 데이터의 확보, 대용량 데이터를 처리하기 위한 인프라의 구축 등 정보 분석에 관심을 가지는 일반 사용자들이 쉽게 해결하지 못하는 어려움이 존재한다. 이 논문에서는 수학적 알고리즘을 기반으로 경제물리학 분야에서 다양하게 응용되고 있는, MST 알고리즘을 활용하기 위한 정보 분석 프로세스를 정의하고 이 프로세스를 수행할 수 있는 분석모듈과 프로토타입을 소개한다. 개발된 프로토타입에서 일반 사용자는 분석에 필요한 주요 파라미터를 선택하고, 서버에서는 Raw 데이터의 전처리 과정을 거쳐 MST를 생성하여 결과를 사용자에게 전송한다.

1. 서론

데이터 분석의 중요성과 이를 통한 새로운 가치 창출은 다양한 분야에서 시도되고 있다. 주식 데이터 분석은 관련 종사자나 전문가, 투자자가 아니더라도 경제적 흐름을 파악하고 이를 활용할 수 있는 중요한 분석 영역이라고 할 수 있다. 특정기업 또는 일반적인 가격변동의 추이 등은 ISP 또는 투자전문기업에서 제공하는 간단한 분석 데이터를 사용자가 확인하는 형식이다. 그러나, 분석기간을 수십년으로 하고, 모든 기업에 대한 분석 그리고 기업간의 관계를 확인하고자 하는 영역으로 본다면, 주식데이터의 확보와 고도화된 알고리즘의 적용, 대용량 데이터를 처리하기 위한 인프라의 구축이 필요하며, 이는 일반적인 사용자가 해결하기 어려운 부분이다. 본 논문에서는 수학적 이론을 기반으로 경제물리학 분야에서 다양하게 연구되고 있는 MST(Minimum Spanning Tree)알고리즘을 적용하여 국내 시장뿐만 아니라 주요 국가의 주식시장을 분석하기 위한 프로세스를 정의하고 모듈화된 프로토타입을 소개한다. 구축된 프로토타입은 일반 사용자가 분석에 필요한 주요 파라미터를 선택한 후 서버로부터 분석결과를 전송받는 모델이므로 주식 데이터 분석의 어려움을 극복하여 일반 사용자들이 활용할 수 있도록 개발되었다. 2장에서는 MST 알고리즘의 기본적인 원리와 주식 데이터에 적용되었을 때의 특징을 설명하고, 3장에서는 MST 알고리즘을 주식 정보 분석에 적용하기 위한 프로세의 정의, Raw 데이터의 수집과 전처리, 개발된 프로토타입의 특징을 설명한다. 4장에서는 결론과 향후과제에 관하여 언급한다.

2. MST를 활용한 주식정보 분석

일반적으로 금융시장은 매우 복잡한 하부구조를 가지는 복잡계(complex system)로 알려져 있으며, 이러한 관점에서 시장을 분석하려는 연구가 경제학자, 수학자, 물리학자들에 의해 이루어지고 있다. 경제학과 물리학의 학제간 연구인 경제물리이론 분야에서는 자산수익률 사이의 상관성 구조를 네트워크 위상구조 (network topology)의 관점에서 파악하려고 하는 MST (Minimum Spanning Tree)라는 방법론이 제기되어 꾸준히 연구가 진행되고 있다[1][2]. 이 알고리즘의 분석에 사용되는 주식수익률은 주가의 로그차분값을 이용한다. t 시점의 데이터를 Y_t 라 하고 이보다 한 시점 이전의 데이터를 Y_{t-1} 이라고 하면, 각각의 자연로그 변환을 취한 것은 $y_t \equiv log(Y_t)$, $y_{t-1} \equiv log(Y_{t-1})$로 정의되며 다음의 사실을 알 수 있다.

$$\begin{aligned} y_t - y_{t-1} &= \log(Y_t) - \log(Y_{t-1}) \\ &= \log(\frac{Y_t}{Y_{t-1}}) \\ &= \log(1 + \frac{Y_t - Y_{t-1}}{Y_{t-1}}) \\ &\cong \frac{Y_t - Y_{t-1}}{Y_{t-1}} \end{aligned}$$

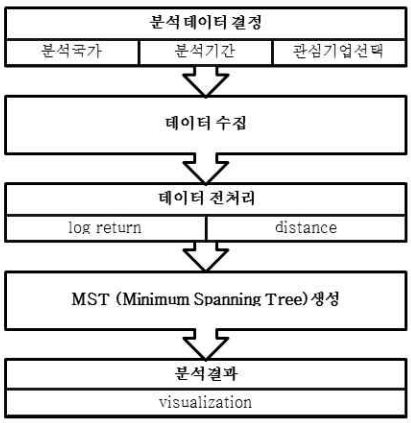
즉, 원래의 데이터에 로그변환을 취한 것의 차분이 바로 원데이터의 변화율을 나타낸다[3]. 대상이 되는 모든 주식 수익률 (n개)에 대해 n×n 의 상관계수행렬을 계산하며, 정의에 의해 상관계수행렬의 각 원소는 -1에서 +1 사이의 값을 가질 것이다. n 개의 주식이 연결되어 있는 거리

공간에서 거리행렬이 주식을 연결하는 MST를 결정할 수 있다. MST는 상관계수행렬로부터의 정보에 기반하여 주식수익률 사이의 상관성을 네트워크 위상구조 (network topology)에서 파악하므로 계수추정치의 절대값에 크게 의존하지 않으면서 리스크를 계량화할 수 있는 효과적인 방식이다. 이러한 MST에서 개별 주식은 일정한 수의 링크(link)를 통해 다른 주식과 연결되는데, 이 수가 해당 주식의 체계적 위험 정도를 나타내는 것으로 간주될 수 있다. 즉, 이 수가 많을수록 시장 전체와 긴밀하게 연결되어 있으며, 체계적 위험이 크다고 할 수 있다. MST의 네트워크에서 멀리 떨어져 있을수록 시장 전체와의 상관성, 즉 체계적 위험은 낮다고 할 수 있다 [1].

3. MST 기반 주식 정보 분석 모듈과 프로토타입

앞장에서 설명한 것처럼 MST 알고리즘을 활용한 주식정보의 분석은 기업간의 관계 분석과 포토폴리오 구성에 가치있는 정보를 제공할 수 있으며, 금융시장 전문가, 수학, 경제학, 물리학 연구자들에 의해 다양한 관점에서의 연구결과들이 생성되고 있다.

그러나, 데이터 확보의 어려움, 데이터 전처리 과정과 분석과정에서의 전문적인 알고리즘의 이해부족, 대용량 데이터를 처리하기 위한 인프라의 부족 등의 이유로 일반 사용자 또는 정보에 관심을 가지고 있는 투자자들이 이러한 기법을 쉽게 활용하는데 어려움을 가지게 된다. 본 논문에서는 MST 알고리즘을 활용한 주식정보의 분석 과정을 모듈화하고, 서버-클라이언트 모델로 구성하여 일반 사용자 또는 연구자들이 활용할 수 있도록 시스템을 구축하는 과정의 프로토타입을 소개한다. 앞장에서 설명한 MST 알고리즘을 적용하기 위해서는 그림 1과 같은 프로세스를 정의할 수 있다. 이러한 각 단계를 모듈화하고 시스템으로 구축한다면, 데이터 수집, 데이터의 전처리, MST의 생성, 분석결과 도출 등의 일련의 과정을 자동화할 수 있다.



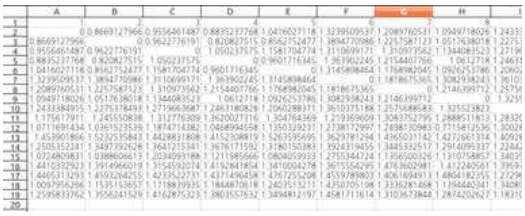
(그림 1) MST를 활용한 주식정보 분석 프로세스

본 논문에서 설명하는 프로토타입은 DATASTREAM[4] 서비스로부터 표 1과 같은 테스트 데이터를 확보하고 Raw 데이터로 활용하였다.

(표 1) 테스트 Data set

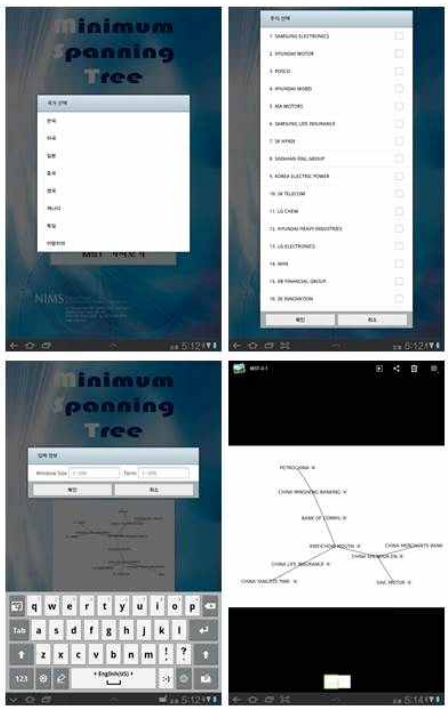
종류	국가	기업수	기간
주식	한국	764	1980.01.01. - 2013.07.31
	미국	533	
	일본	180	
	중국	300	
	영국	150	
	독일	60	
	프랑스	100	
	이탈리아	80	

분석모듈은 사용자가 선택하는 분석국가, 기간, 관심기업에 따라 Raw 데이터를 활용하여 log return, distance 계산(그림 2), MST 생성, 분석결과(관계도)를 생성하게 되고 이 결과를 통해 사용자는 주식시장을 새로운 시각에서 이해하고 정보를 활용할 수 있다.



(그림 2) MST를 생성하기 위한 데이터 전처리

이러한 자동화된 모듈로 구성된 프로토타입은 서버-클라이언트 모델[6]로 구성되어, 사용자는 분석을 위한 주요 파라미터를 선택하고, 서버에서는 전송받은 분석 파라미터와 Raw 데이터를 활용하여 분석을 수행하고 결과를 사용자에게 전달하는 역할을 수행하게 된다.



(그림 3) MST 분석 프로토타입 (screen shot)

프로토타입은 Android SDK [5]로 개발되어 모바일 디바이스(스마트폰, 태블릿)에서 실행되도록 구성되어 있다. 그림 3은 사용자의 모바일 디바이스에서 분석 파라미터(국가, 기업, 분석기간) 선택 UI와 사용자에게 전송되는 분석결과(MST)의 예를 보여주고 있다.

4. 결론과 향후과제

본 논문에서는 대용량의 주식데이터로부터 기업간의 관계를 파악하고, 포토폴리오 구성에도 활용할 수 있어 많은 연구가 진행되고 있는 MST 알고리즘 기반의 주식데이터 분석을 일반 사용자들도 활용할 수 있도록 모듈화하고 프로토타입으로 개발하였다. 전문적인 지식과 인프라를 가지고 있지 않더라도 다양한 데이터 분석이 가능하도록 하는 것은 데이터의 편중된 활용을 방지하기 위해서도 중요한 일이다. 본 논문에서 분석과정을 모듈화 한 이유는 향후 개발되는 다양한 알고리즘이 쉽게 적용되게 하기 위해서이다. 향후 과제는 Raw 데이터 수집의 자동화, 분석모듈의 다양화, 분석결과 표현의 고도화 등을 확보하여 시스템의 완성도를 높이는 것이다.

참고문헌

[1] 조하연, 이승국, “MST 기법을 이용한 주식시장의 상관성 구조와 체계적 위험에 관한 연구”,한국금융학회,2004-04.

[2] L. Sandoval Jr., “Pruning a mimimum spanning tree”, Physica A, vol. 391, issue 8, pp. 2678-2711, April 2012.

[3] 한창호,“시계열 데이터에 로그변환을 취하는 이유”,쾐트글로벌 금융공학 길잡이, 2009.

[4] THOMSON REUTERS DATASTREAM; https://forms.thomsonreuters.com/datastream/

[5] Android SDK; http://developer.android.com/sdk/index.html

[6] 허준, 권오규, 오준상, 장영재, “모바일 앱을 활용한 사용자 중심 데이터 분석 프레임워크”, 정보및제어학회 (CICS 13) 2013.