Vol.3, No.10 pp.441~450 pISSN: 2287-5905 http://dx.doi.org/10.3745/KTSDE.2014.3.10.441

The Stock Portfolio Recommendation System based on the Correlation between the Stock Message Boards and the Stock Market

Yun-Jung Lee† · Gun-Woo Kim†† · Gyun Woo†††

ABSTRACT

The stock market is constantly changing and sometimes the stock prices unaccountably plummet or surge. So, the stock market is recognized as a complex system and the change on the stock prices is unpredictable. Recently, many researchers try to understand the stock market as the network among individual stocks and to find a clue about the change of the stock prices from big data being created in real time from Internet. We focus on the correlation between the stock prices and the human interactions in Internet especially in the stock message boards. To uncover this correlation, we collected and investigated the articles concerning with 57 target companies, members of KOSPI200. From the analysis result, we found that there is no significant correlation between the stock prices and the article volume, but the strength of correlation between the article volume and the stock prices is relevant to the stock return. We propose a new method for recommending stock portfolio base on the result of our analysis. According to the simulated investment test using the article data from the stock message boards in 'Daum' portal site, the returns of our portfolio is about 1.55% per month, which is about 0.72% and 1.21% higher than that of the Markowitz's efficient portfolio and that of the KOSPI average respectively. Also, the case using the data from 'Naver' portal site, the stock returns of our proposed portfolio is about 0.90%, which is 0.35%, 0.40%, and 0.58% higher than those of our previous portfolio, Markowitz's efficient portfolio, and KOSPI average respectively. This study presents that collective human behavior on Internet stock message board can be much helpful to understand the stock market and the correlation between the stock price and the collective human behavior can be used to invest in stocks.

Keywords : Stock Market, Stock Market Volatility, Stock Network, Stock Portfolio, Stock Message Board

인터넷 주식 토론방 게시물과 주식시장의 상관관계 분석을 통한 투자 종목 선정 시스템

이 윤 정† · 김 건 우†† · 우 균†††

요 약

주식시장은 항상 변하며 특별한 이유 없이도 주가가 급락하거나 급등하는 현상도 나타난다. 그러므로 주식시장은 복잡계로 인식되고 있으며, 주가의 변화는 예측하기 어렵다. 최근에 많은 연구자는 주식시장을 개별 주식 간의 네트워크로 간주하고 그것을 이해하려고 하며, 인터넷에서 실시간으로 생성되는 빅데이터를 통해 주가의 변화를 밝히려고 노력하고 있다. 우리는 주가와 인터넷 특히 주식토론방에 나타나는 사람들의 반응 간의 상관관계에 주목한다. 이 상관관계를 밝히기 위해서 KOSPI200에 속한 회사 중 57개 회사와 관련 있는 게시물들을 수집하고 분석하였다. 분석 결과에 따르면, 개별 주가와 게시물 수 사이에는 특별한 상관관계가 나타나지 않았지만, 주가와 게시물 수의 상관관계가 주식 수익률과 관계가 있는 것으로 나타났다. 우리는 이 분석결과를 기반으로 주식투자 포트폴리오를 추천하는 새로운 방법을 제안한다. '다음' 포털의 주식토론방 데이터를 이용한 모의 투자 실험 결과에서, '다음' 주식토론방 데이터를 사용한 경우 제안 방법으로 구성한 주식 포트폴리오의 월평균 수익률은 약 1.55%로 마코위츠의 효율적 포트폴리오의 수익률보다 약 0.72% 높으며, 코스피 평균 수익률보다 약 1.21% 높게 나타났다. 또한 '네이버' 주식토론방 데이터를 사용한 경우는 모의 투자 수익률이 약 0.90%로 기존 방법과 마코위츠 효율적 포트폴리오와 코스피 평균 수익률보다 각각 0.35%와 0.40%, 0.58% 높게 나타났다. 이 연구는 인터넷 주식토론방에 나타난 사람들의 집단적인 행위는 주식시장을 이해하는 데 많은 도움을 줄 수 있으며, 주가와 사람들의 집단행위 사이의 상관관계가 주식투자에 활용될 수 있음을 제시하였다.

키워드 : 주식시장, 주식시장 변동성, 주식 네트워크, 주식 포트폴리오, 주식토론방

※ 본 논문은 2013년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2013S1A5B6053791).
† 정 회 원 : BK21PLUS IT기반융합산업창의인력양성사업단 박사후연구원
†† 준 회 원 : 부산대학교 전기컴퓨터공학과 석사과정
††† 종신회원 : 부산대학교 전기컴퓨터공학과/LG 스마트제어센터 교수

Manuscript Received : June 27, 2014
First Revision : August 27, 2014
Accepted : September 9, 2014
* Corresponding Author : Gyun Woo(woogyun@pusan.ac.kr)

1. 서 론

주식시장에서 주가에 영향을 주는 요인으로는 유가나 환율과 같이 직접적인 측정이 가능한 요인에서부터 시장 상황이나 외부 위험과 같은 눈에 보이지 않는 요인들까지 수없이 많다. 따라서 주식시장은 시간에 따라 계속 변하고 특별한 이유 없이 주가가 급등하거나 급락하는 사건들이 발생하기도 하여 주가를 예측하는 것은 매우 어려운 일이다.

최근에는 주식시장을 이해하기 위해서 금융 관련 분야뿐만 아니라 통계나 전산 등 다양한 분야에서 주식시장을 복잡계로 인식하고 복잡계에서 나타나는 특징들을 찾아내고 분석하는 연구도 이루어지고 있다[1-3]. Mantegna는 주식시장을 구성하는 개별 종목들의 상관관계를 기반으로 하여 주식시장을 최소신장트리(Minimum Spanning Tree, MST)로 나타내어 전체 주식시장의 위상 구조를 이해하려고 시도하였다[4].

그뿐만 아니라 최근에는 트위터나 인터넷 검색, 인터넷 뉴스 등과 같은 온라인 매체를 분석함으로써 주식시장과의 상관관계를 분석하려는 연구가 진행되고 있다[5-8]. Preis와 동료들은 금융과 관련된 검색어의 검색량 변화가 미래의 주가 변동의 조기 신호로 해석될 수 있다고 제시하였다[7]. 이들은 98개의 검색어를 이용한 모의 투자 실험을 통해 구글 검색 데이터가 현재의 경제 상황을 반영할 뿐만 아니라 미래의 경제 활동에 관한 추세를 파악하는 데 사용될 수 있음을 발견하였다. 이외에도 인터넷을 통해 실시간으로 생성되는 빅데이터를 분석함으로써 주식시장의 변화를 이해하려는 연구가 계속되고 있다.

이 논문에서는 주식시장과 인터넷 주식토론방에 나타난 사람들의 집단행동과의 상관관계를 분석한다. 또한, 분석 결과와 주식 네트워크의 특성을 이용하여 새로운 주식 포트폴리오 구성 알고리즘을 제안한다. 제안 방법의 효율성을 보이기 위해 실제 KOSPI200을 구성하는 57개 회사의 주식을 대상으로 포트폴리오를 구성하고, 제안 포트폴리오의 수익률과 코스피 수익률을 비교하여 제안 방법의 효율성을 평가한다.

논문의 구성은 다음과 같다. 2절에서는 주식시장과 인터넷 데이터와의 관계를 분석한 선행 연구들을 살펴본다. 3절에서는 실제 주식 데이터를 이용하여 얻은 주식 네트워크를 분석하고, 4절에서 포트폴리오 구성 방법에 대해서 설명한다. 그리고 5절에서는 모의 투자 실험을 통해 제안 포트폴리오의 효율성을 평가하고, 6절에서 결론을 맺는다.

2. 관련 연구

주식시장에서는 국내외적으로 발생하는 사건으로 인하여 예상치 못한 급락이 발생하는 등 전형적인 복잡계 현상이 나타나고 있다. 이것은 주식 간 혹은 시장 간의 눈에 보이지 않는 연결로 정보에 대한 직접 또는 간접적인 영향 경로가 형성되어 있기 때문이다[9]. 1999년 Mantegna의 연구를 시작으로 주식 간 혹은 시장 간의 상호 관계를 네트워크 방법으로 이해하려는 연구가 주로 경제물리학 분야에서 진행되었다. Mantegna는 주식 사이의 복잡한 상호작용을 바탕으로 주식들이 동질적 집단을 형성하는 집단화 과정을 주식 네트워크로 시각화하였다[4]. 이 그래프에서 각 노드는 주식 종목이고, 노드 간 연결 거리는 두 주식 사이의 상관관계로 구해진다.

Fig. 1은 다우존스지수를 계산하는 데 사용되는 종목 중 30개 종목을 연결한 최소신장트리를 보여준다. 그림에서 각각의 노드는 회사의 주식을 나타내고, 각 노드의 심볼은 회사명을 나타내는 약어인 티커심볼(ticker symbol)로 'GE'는 'General Electric Company'를 가리킨다. 각 노드 간의 연결 거리는 두 주식의 상관관계를 이용하여 구한 것으로 거리가 짧을수록 두 주식 간의 상관관계가 높음을 의미한다. 이 그래프에서는 주식 'CHV'와 주식 'TX' 사이의 거리가 가장 짧게 나타났다.

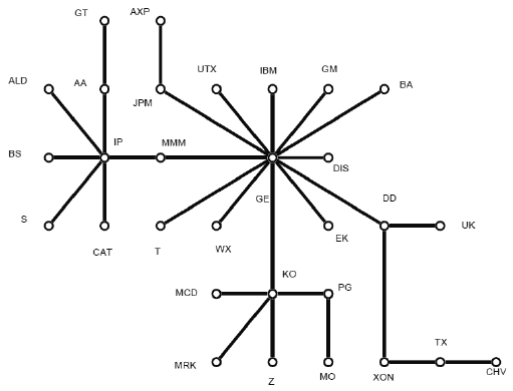


Fig. 1. The minimal spanning tree connecting the 30 stocks used to compute the DJIA[4]

Mantegna가 주식시장 분석을 위해 네트워크 방법을 도입한 후로 주식시장 네트워크가 경제적인 관점에서 신뢰할 수 있는 것인지에 대한 다양한 분석과 검증이 시도되었다. Onnela와 동료들은 주식시장 그래프의 시간에 따른 변화를 관찰하였다[1]. 1980년 1월 2일부터 1999년 12월 31일까지 약 20년 동안의 뉴욕 증시에서 거래되는 주식 종목들의 주가를 대상으로 4년 단위로 연속적인 주식시장의 최소신장트리를 구성한 결과 기본적인 트리의 위상구조는 시간에 따라 큰 변화는 없었으나 1987년 10월 19일인 '블랙 먼데이(Black Monday)'와 같은 주식 대폭락기에는 트리의 평균경로 길이가 감소한 것으로 나타났다.

또한, 이 연구에서는 주식시장의 최소신장트리를 포트폴리오 분석에도 적용하였다. 결과에 따르면 마코위츠(Markowitz)에 의하여 제안된 최적화 함수로 구성한 효율적 포트폴리오를 분석한 결과 포트폴리오 구성 주식들이 주식시장 그래프의 단말 노드에 위치하는 경향이 있는 것으로 나타났다.

지금까지 미국 증시에 대한 분석뿐만 아니라 여러 나라의 주식시장 네트워크에서 나타나는 특성을 분석하는 연구도 활발히 이루어졌다. 엄철준과 동료들은 한국 주식시장에서 관찰 가능한 주식 연결 구조가 어떤 방식으로 형성되는지 연구하였다[10]. 한국 주식시장 그래프에서 주식 대부분은 다른 주식들과의 연결 수가 1개 혹은 2개로 적지만 일부 몇몇 주식들의 연결 수는 매우 많은 것으로 나타났으며, 즉 주식 네트워크에서 개별주식의 연결선 수 분포는 거듭제곱 분포를 따른다는 공통적인 특성이 관찰되었다. 또한, 많은 연결 관계가 형성된 중심 주식(hub stock)일수록 시장지수와 관련성이 높은 것으로 나타났다.

이와 비슷한 연구로 허화와 동료들은 한국 주식시장의 주식 네트워크와 주식 수익률이 보이는 상관관계에 대해 분석하였다[9]. 이 연구에서는 1980년 1월부터 2003년 5월까지 23년 동안의 연속적인 일별 주식 가격 정보를 갖는 197개의 주식을 대상으로 주식 네트워크와 최적화 함수에 따라 도출된 효율적 포트폴리오를 비교하였다. 포트폴리오에 속한 주식들은 대부분 주식 네트워크에서 하나 또는 둘 정도의 적은 연결선을 가지고 있으며 그래프의 외곽에 위치하는 경향이 나타난다는 사실이 확인되었다. 이 결과는 앞서 설명한 Onnela의 연구와도 일치한다.

이외에도 중국이나 러시아, 호주 등 여러 나라의 주식 네트워크 특성을 분석한 연구들이 많이 발표되었다[11-13]. 나라마다 주식시장의 규모나 성숙도 등에 따라 차이는 있으나 주식 사이의 연결선 분포나 업종 간 구분 등과 같은 주식 네트워크에서 나타나는 공통적인 특성들이 관찰되었다. 따라서 주식 네트워크가 실제 주식시장에서 나타나는 상관관계 등을 의미 있는 수준으로 반영하고 있으며, 이로써 네트워크 방법론이 주식시장을 이해하는 데 도움이 되는 방법임을 알 수 있다.

최근에는 트위터나 인터넷 검색, 인터넷 뉴스 등과 같은 온라인 매체에서 생성되는 빅데이터를 이용하여 주식시장과의 상관관계를 분석하려는 연구가 진행되고 있다[6, 8, 14]. J. Bollen와 동료들은 대규모 트위터 메시지 데이터에 나타난 대중의 분위기가 다우존스지수와 관계가 있는지를 분석하였다[5]. 그들은 매일 트위터 메시지에 대해, OpinionFinder를 이용하여 긍정과 부정의 두 가지 감정을 측정하고, GPOMS (Google-Profile of Mood State)를 이용하여 평온과 놀람을 포함한 6가지 감정 상태를 측정하였다. 그 결과 이런 대중들의 분위기가 약 3.4일 후의 다우존스지수와 부합하는 것으로 나타났다.

또 다른 연구로 Preis와 동료들은 인터넷에서 사람들의 상호작용으로 발생하는 새로운 거대한 데이터가 거대한 시장의 움직임에서 시장 참여자들의 행위를 반영한다고 제시하였다[7]. 그들은 주식시장과 연관된 98개의 검색어의 성능을 분석하였다. 2004년 1월부터 2011년 2월까지 검색 기록을 분석한 결과, 주식시장이 폭락하기 전에 구글에서 금융 관련 검색어의 검색량이 증가한다는 것을 발견하였다. 이 결과를 바탕으로 그들은 금융에 관련된 단어 검색량의 변화가 주식시장 움직임의 '조기경보' 역할을 할 수 있을 것으로 주장하였고, 그것을 이용하여 적절한 매매 전략을 세울 수 있다고 제시하였다.

이렇듯 최근에는 인터넷을 매개로 하는 다양한 데이터를 활용하여 주식시장의 변화를 이해하려는 연구가 진행되고 있다. 현재는 트위터 메시지나 포털 검색 데이터를 이용하는 연구가 많이 발표되고 있지만, 주식이나 금융 관련 인터넷 뉴스나 주식토론방 등도 주식시장을 이해하는 데 활용할 수 있을 것으로 생각하며, 오히려 단순한 검색 정보보다 더 정확한 경향 정보를 줄 수 있을 것으로 생각한다.

3. 인터넷 주식토론방 사용자의 집단행위 분석

인터넷 주식토론방은 주식이나 금융 관련 이슈에 관해 서로의 의견을 주고받는 공간이다. 주식토론방의 게시물을 읽음으로써 사용자들은 자신이 관심 있는 주식이나 회사에 대한 정보나 시장 분위기를 파악할 수 있다. 이 절에서는 주가와 주식토론방에서 나타나는 사람들의 집단적인 반응 사이에 어떠한 상관관계가 있는지를 분석한다.

3.1 주식토론방 데이터

주식토론방 사용자들의 집단행위를 분석하기 위해서 국내 유명 포털 사이트인 '다음(Daum)'에서 제공하는 주식토론방에 등록된 게시물을 수집하였다. '다음' 포털에서는 코스피(KOSPI) 전 종목에 대해서 게시판을 운영하고 있으며 회사별로 게시판이 따로 마련되어 있어 게시물을 읽지 않고도 회사별로 구분할 수 있다. Table 1은 '다음' 주식토론방에 개설된 코스피 종목 게시판의 현황을 보여준다.

Table 1. The basic statistics of the stock message board of 'Daum'

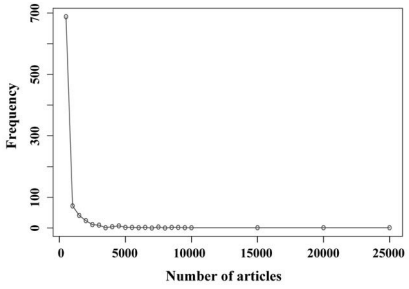
Board	Count	883
	Start date	2002.06.24
Number of article	Total	490,862
	Mean	555.90
	Stdev.	1550.85
Articles per company	Daily mean	0.25
	Max	23,507
	Min	0

주식토론방에는 코스피를 구성하는 883개 회사의 토론방이 개설되어 있으며 2002년 7월 24일에 첫 번째 게시물이 등록되었다. 2014년 4월 기준으로 약 490,862개의 게시물이 등록되었으며, 평균적으로 회사별 약 550개의 게시물이 등록되었다고 할 수 있다.

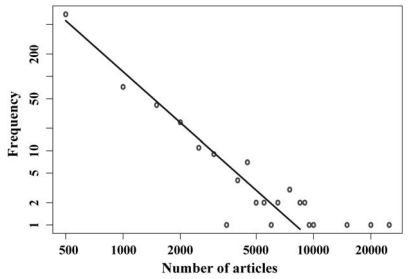
회사별 게시물 등록 현황을 살펴보면 가장 많은 게시물이 등록된 곳은 'SK 하이닉스'로 23,507개의 게시물이 등록되었다. 반면에 6개의 회사는 게시판 개설 이후로 하나의 게시

물도 등록되지 않았다. 회사별 하루 평균 게시물 수는 0.25개로, 아직 주식토론방 사용이 모든 회사에서 활발한 것은 아님을 알 수 있다.

Fig. 2는 '다음' 주식토론방에 개설된 883개 게시판의 게시물 분포를 보여준다.



(A) The distribution of 490,862 articles obtained from 'Daum'



(B) The distribution of the articles (log-log scale)

Fig. 2. The distribution of the articles concerning target 883 companies

Fig. 2A에서 x축은 각 게시판에 등록된 게시물 수를 나타내고, y축은 게시판 수, 즉 회사 수를 나타낸다. Fig. 2B는 Fig. 2A와 같은 그래프를 로그 축으로 나타낸 것이다. 그림에서 회사별로 등록된 게시물 수는 거듭제곱 분포로 나타남을 알 수 있다. 이것은 주식토론방 사용이 활발한 회사와 그렇지 않은 회사 사이에 편차가 크다는 것을 의미한다.

3.2 주식시장과 주식토론방 사용자들의 집단행위 사이의 상관관계

주식토론방에서 일어나는 사람들의 집단행위와 주가 변동 사이의 상관관계를 살펴보기 위해서 각 회사의 주가와 게시물 수의 변화를 살펴보았다. 한 회사의 주가 변동과 게시물 수의 상관관계를 계산하기 위해서는 관찰 기간 동안 지속적으로 게시물이 등록되어야 한다. 따라서 이 논문에서는 분석 대상 주식을 KOSPI200에 속한 회사 중에서 하루 평균 약 0.5개 이상의 게시물이 등록된 회사의 주식으로 제한하였고, 관찰 기간은 2008년 1월 1일부터 2013년 12월 31일까지 총 6년으로 설정하였다. 분석 대상 회사의 게시판에서 수집된 게시물 데이터는 Table 2와 같다.

Table 2. The state of the target stock message boards where were posted more than 0.5 articles daily (from January 1 2008 to December 31 2013)

Company	Count	57
Number of articles	Total	138,618
	Daily Mean	1.63

분석 대상 회사는 모두 57개로 조사되었으며, 6년 동안 이들 회사의 주식과 관련된 게시물들은 총 138,618개로 회사별 하루 평균 약 1.63개의 게시물이 등록된 것으로 나타났다. 57개의 회사 중 게시물이 가장 많은 것은 'SK 하이닉스'사로 6년 동안 14,117개로 하루 평균 약 9.44개의 게시물이 등록된 것으로 나타났다. Fig. 3은 'SK 하이닉스'사의 월별 게시물 수를 보여준다.

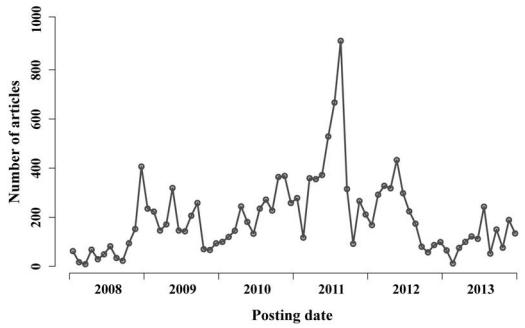


Fig. 3. Monthly changes of the article volume posted on the 'SK hynics' board

게시물이 가장 많이 생성된 기간은 2011년 8월로 한 달 동안 918개의 게시물이 주식토론방에 등록되었다. 이 기간에 많은 게시물이 등록된 이유는 'SK 하이닉스' 회사의 전신인 '하이닉스반도체'의 매각과 관련해 많은 사람이 반응한 것으로 추정된다. 이와는 반대로 2008년 3월에는 9개의 게시물밖에 등록되지 않았다. 이처럼 게시물 수의 변화 그래프에서도 볼 수 있듯이 기간별로 등록되는 게시물 수의 변동이 큼을 알 수 있다.

주식토론방은 자신의 관심 주식 종목에 관한 의견들을 주고받는 게시판이다. 따라서 주가에 영향을 줄만한 사건이 발생하거나 주가의 변동에 따라 게시물의 내용이나 양도 영향을 받을 것이다.

Fig. 4는 'SK 하이닉스'사의 주가와 게시물 수의 시계열 데이터를 나타낸 것으로, 두 값을 정규화하여 비교하였다. 막대그래프와 실선 그래프는 각각 z-score로 변환한 게시물 수와 주가를 나타낸다. 그림에서 전반적으로는 주가가 하락하는 시기에 게시물량이 증가하는 경향이 나타나며, 특히 주가가 급등하거나 급락하는 등 주가 변동이 큰 시기에 게시물의 수도 함께 증가하는 것으로 보인다. 특히 주식토론방 이용자들은 주가가 상승할 때보다는 하락하는 시기에 더 민감하게 반응하는 경향이 있는 것으로 나타났다.

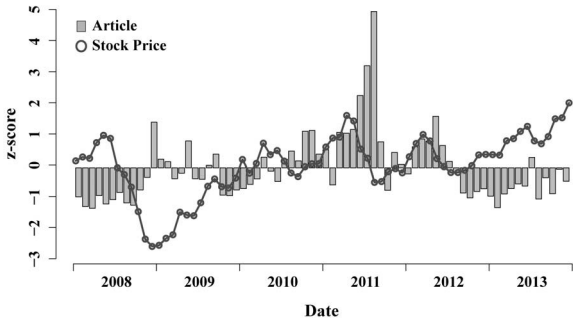


Fig. 4. Changes of the stock price (piecewise line) comparing with the article volume (bar chart) of 'SK hynics'

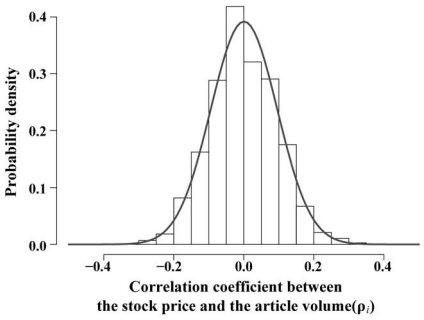
이와 같은 주가와 주식토론방 게시물의 상관관계가 어느 정도인지를 살펴보고, 회사마다 상관관계 정도가 비슷한지를 살펴보기 위해 회사별로 주가와 토론방 게시물 수의 상관관계를 분석하였다. 분석을 위해 주가와 게시물 수의 상관계수 ρ_i는 Equation (1)을 이용하여 6개월 간격으로 측정하였다.

$$\rho_i = \frac{\Sigma_i (s_i - \overline{s})(a_i - \overline{a})}{\sqrt{\Sigma_i (s_i - \overline{s})^2}\sqrt{\Sigma_i (a_i - \overline{a})^2}} \quad (1)$$

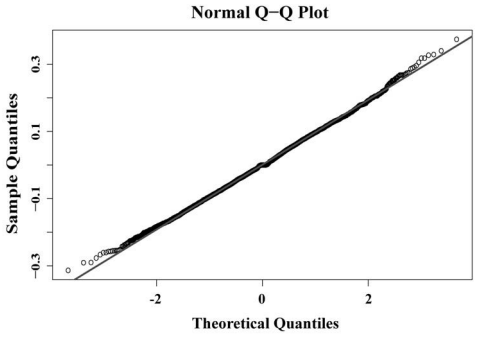
여기서 s_i와 a_i는 각각 회사 i의 주식 가격과 주식토론방에 등록된 회사 i에 관한 게시물 수를 나타낸다. 그리고 $\overline{s}$와 $\overline{a}$는 s_i와 a_i의 평균을 가리킨다.

Fig. 5는 Table 2의 57개 회사에 관한 주가와 게시물 수 사이의 상관계수 분포를 보여준다. Fig. 5A는 주가와 게시물 수 사이의 상관계수 데이터의 히스토그램으로, 그림에서 붉은색 그래프는 상관계수 데이터와 동일한 평균과 표준편차를 갖는 정규분포 곡선을 가리킨다. 그림에서 상관계수 데이터가 정규분포에 가까운 것을 알 수 있다. 분석 데이터의 평균은 약 0.0006이고, 표준편차는 약 0.0949로 계산되었다.

Fig. 5B는 상관계수 데이터를 Q-Q 플롯으로 표현한 것이다. Q-Q 플롯은 데이터의 분포가 정규분포에 얼마나 가까운지 직관적으로 표현한 그래프로, 그림에서 붉은색 직선이 데이터와 같은 평균과 표준편차를 가지는 정규분포를 가리킨다. 따라서 데이터가 붉은색 직선에 가까울수록 정규분포에 가깝다는 것을 의미한다. 그림에서 주가와 게시물 수의 상관계수 데이터는 거의 정규분포에 가깝지만 직선의 양 끝단에서 정규분포와는 약간의 차이가 발생하는 것을 볼 수 있다. 이것은 앞서 살펴본 것과 같이 대부분의 시기에는 주가와 게시물 수 사이에 특별한 상관관계가 없다가 주가가 하락하거나 변동 폭이 큰 특정 시기에 상관관계가 높아지는 것과 같은 맥락으로 이해할 수 있다.



(A) Histogram of correlation data and normal distribution curve



(B) Q-Q plot

Fig. 5. Distribution of the correlation coefficients between the stock prices and the article volume (measuring interval: six month)

이렇듯 주가와 게시물 수 사이에 직접적인 상관관계가 나타나지는 않지만, 주가의 변동 폭이 큰 시기에 사람들의 반응도 함께 커지는 경향이 있음을 앞서 살펴보았다. 따라서 주가의 변동 폭은 주식수익률과 어느 정도 관련이 있다고 할 수 있다.

다음으로 주가와 게시물 수의 상관관계와 주식수익률과의 관계를 분석하였다. Fig. 6은 주가와 게시물 간의 상관계수(ρ_i)와 해당 기간의 주식수익률 분포를 보여준다. 그림에서 가로축은 주가와 게시물의 상관계수를 나타내고, 세로축은 주식수익률을 나타낸다. 상관계수와 주식수익률은 6개월 단위로 측정하였다. 분석한 결과 상관계수 값과 수익률 사이에 약한 선형관계가 나타나는 것을 볼 수 있다. Fig. 6에서 주목할만한 것은 가로축(ρ_i)의 양 끝단으로 갈수록 수익률이 높거나 낮게 치우치는 것이다. 즉, 주가와 게시물 사이에 강한 양의 상관관계가 나타나는 경우 양의 수익률을 보일 확률이 높으며, 강한 음의 상관관계가 있는 경우 수익률도 음일 확률이 높은 것으로 추정된다.

앞의 두 분석 결과를 종합해보면, 주가와 토론방 게시물 간에 직접적인 상관관계는 없으나 주가가 급락하거나 급등하는 시기에는 토론방 사용자들의 반응이 증가하며, 이러한 주가와 사용자들의 반응 사이의 상관관계가 높을수록 양의 수익률을 보일 확률이 높다고 할 수 있다.

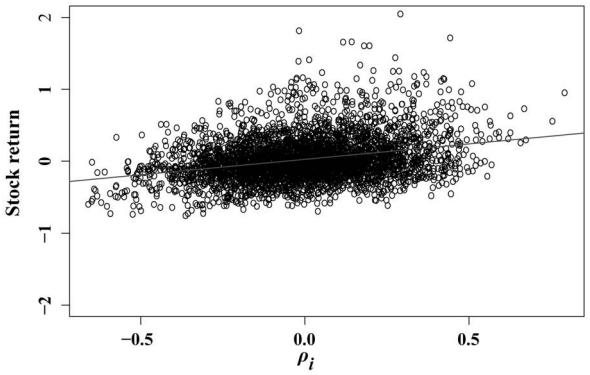


Fig. 6. Distribution of the stock returns depending on the correlation coefficient between the stock prices and the article volume

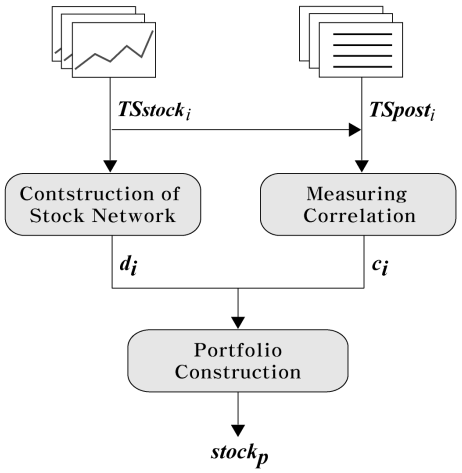


Fig. 7. Flowchart of a proposed method for recommending a portfolio

4. 투자 종목 추천 시스템

이 논문에서는 시간에 따라 변하는 주식시장에서 주가와 게시물의 상관관계를 이용하여 투자 종목을 추천하는 방법을 제안한다. 우리는 기존에 주식시장의 네트워크 특성을 이용하여 주식 포트폴리오를 구성하는 알고리즘을 제안하였다[15]. 기존 방법에서는 효율적 포트폴리오 구성 종목들이 주식 네트워크에서 외곽에 위치하는 경향이 있다는 연구 결과를 바탕으로 하고 있다[9]. 이 방법에서는 K-means 알고리즘으로 주식들을 클러스터링하고 각 클러스터에서 차수가 1인 종목을 하나씩 선택한다. 따라서 K-means 알고리즘의 특성상 시뮬레이션을 수행할 때마다 클러스터의 결과가 달라질 수 있다. 또한, 차수가 1인 종목이 여러 개일 경우 무작위로 선택하므로 같은 기간에도 실행할 때마다 포트폴리오 구성 결과가 달라질 수 있다.

제안 방법에서는 기존 방법과는 달리 주식들을 클러스터링하지 않고 주가와 게시물 간의 상관관계와 주식 네트워크에서 노드의 차수 정보를 이용하여 포트폴리오 구성 종목을 선택한다. 포트폴리오 구성을 위해 주가와 게시물 수의 상관계수가 임계치 이상이고, 주식 네트워크에서 노드의 차수가 1인 종목을 모두 선택하므로 시뮬레이션 수행 시 결과 포트폴리오는 변하지 않는다. 제안 방법의 알고리즘은 Fig. 7과 같다.

Fig. 7에서 $stock_i$와 $post_i$는 회사 i의 주가와 주식토론방에 등록된 게시물의 시계열 자료이다. 제안 방법은 크게 두 부분으로 구성된다. 먼저 주식 네트워크 구성 모듈에서는 개별 회사 간의 주가 시계열 데이터의 상관관계를 측정하여 주식 네트워크를 구성한다. 이때 주식 네트워크는 개별 주식을 노드로 하고 주식 간 주가 상관계수를 에지 가중치로 하는 최소신장트리로 구성된다. 최소신장트리에서 노드 간 거리는 두 노드를 연결하는 에지 가중치가 높을수록 두 노드의 거리는 짧아지며, 두 노드 i와 j 간 거리 $dist(i,j)$는 Equation (2)와 같이 계산된다.

$$dist(i,j) = \sqrt{2 \bullet (1 - c_{ij})} \tag{2}$$

여기서 c_{ij}는 회사 i와 j의 주가 데이터의 피어슨 상관계수를 가리킨다. Fig. 8은 57개 회사의 2008년 1월 3일부터 2013년 12월 31일까지 6년 동안 주가 데이터를 이용하여 구성한 주식 네트워크를 보여준다. Fig. 8에서 각 노드는 업종별로 모양과 색깔을 구별하였으며, 노드의 라벨은 해당 회사의 주식 종목 코드를 가리킨다. 이렇게 주식 네트워크를 구성함으로써 각 노드의 차수 d_i를 구할 수 있다.

다음으로 주가 시계열 자료와 주식토론방에서 해당 종목에 관련된 게시물의 시계열 자료를 이용하여 주가와 게시물 간의 상관계수 ρ_i를 계산한다. 이때 ρ_i는 Equation (1)을 이용하여 계산할 수 있다. 마지막으로 각 회사의 주가와 게시물 상관관계 ρ_i가 임계치 이상이고, 주식 네트워크에서 노드 차수 d_i가 1인 종목이 최종 투자 종목 $stock_p$로 선택된다.

Fig. 8. Constructed stock market with 57 selected stocks (date: from January 1 2008 to December 31 2013)

5. 실험 결과

제안 방법의 효율성을 보이기 위해 실제 주식들을 대상으로 하여 모의 투자 실험을 수행하고 수익률을 살펴보았다. 먼저 실험을 위해서 3절에서 설명한 것처럼 KOSPI200을 구성하는 종목 중 하루 평균 약 0.5개 이상의 게시물이 등록된 57개 회사의 주식을 선정하고, 2008년 1월부터 2013년 12월까지 약 6년 동안의 일별 주가와 같은 기간 '다음'과 '네이버' 포털에서 제공하는 주식토론방 게시물 데이터를 실험 데이터로 사용하였다.

실험은 다음과 같이 구성된다. 실험 데이터를 이용해 직전 1개월의 데이터를 분석하여 다음 1개월의 투자 포트폴리오를 구성하고 수익률을 측정하였다. 포트폴리오 구성을 위해서 주가와 게시물 상관계수가 임계치 이상인 종목 중에서 차수가 1인 주식을 최종 투자 종목으로 선택한다. 실험에서 임계치는 0.3으로 설정하였다. 이때 임계치 이상인 종목 중에서 차수가 1인 종목이 없는 경우에는 차수가 가장 적은 종목을 선택한다. 모의 투자 실험에서는 포트폴리오에 포함된 모든 종목이 같은 비율로 투자된다고 가정한다. 따라서 포트폴리오가 총 n개의 종목으로 구성되었고 전체 투자금을 N으로 가정한다면 각 주식별 투자금액은 n/N이 된다.

제안 방법으로 구성한 포트폴리오의 수익률 분석을 위해 포트폴리오 구성은 1개월 단위로 하고, 1개월씩 이동시켜 모의 투자 실험을 수행하였다. 또한, 제안 방법으로 구성한 포트폴리오가 얼마나 효율적인지를 보이기 위해 주식 네트워크 특성만 고려한 기존 방법과 마코위츠의 효율적 포트폴리오 구성 알고리즘을 이용하여 구성한 포트폴리오의 수익률, 같은 기간의 KOSPI200 평균 수익률을 각각 비교하였다.

먼저 '다음' 포털의 주식토론방 데이터를 이용해서 포트폴리오를 구성하고 모의 투자 실험을 수행하였다. 각각의 수익률은 Table 3에 정리되어 있으며, 지면 관계상 전체 수익률 결과의 일부만 나타내었다. 모의 투자 실험에서 실험 기간의 1개월 수익률 평균은 약 1.55%로, 기존 방법보다는 약 0.67%, 마코위츠 방법보다는 약 0.72%, 코스피 평균 수익률보다는 약 1.21% 높게 나타났다. 실험 결과에서 알 수 있듯이 모의 투자 실험 결과, 제안 방법으로 구성한 포트폴리오가 비교 포트폴리오들의 수익률보다 높게 나타났다.

모의 투자 동안 누적 투자 수익률을 분석한 결과 포트폴리오별 누적 수익률은 Fig. 9와 같이 나타났다. Fig. 9에서 포트폴리오별 누적 투자 수익률의 등락은 투자 기간 중 거의 비슷한 패턴으로 나타나고 있다. 비교 포트폴리오 중에서 제안 방법으로 구성한 포트폴리오가 모의실험 기간 중 거의 전 구간에서 높은 수익률을 보이는 것으로 나타났다.

다음으로 '네이버' 포털에서 제공하는 주식토론방의 데이터를 이용하여 포트폴리오를 구성하고 같은 방법으로 모의 투자를 수행하였다. '네이버' 주식토론방의 경우 '다음' 주식토론방과는 달리 현재 날짜에서 이전 1년 동안의 게시물 데이터만 제공하고 있다.

Table 3. The simulated investment earnings rate of our portfolio and three comparison portfolios (The proposed portfolio has been constructed using the article data of 'Daum.')

Period	Monthly average rate of return (%)			
	Proposed portfolio	Previous portfolio	Markowitz portfolio	KOSPI200
2008-04	3.89	8.83	6.77	7.24
2008-08	-3.50	-5.30	-7.42	-6.32
2008-12	20.23	9.73	3.34	6.22
2009-04	4.02	15.58	5.14	11.03
2009-08	3.14	3.78	10.24	1.72
2009-12	6.97	8.43	4.79	7.20
2010-04	3.95	3.18	3.34	1.30
2010-08	7.11	3.38	4.00	-2.22
2010-12	10.47	9.06	8.33	6.31
2011-04	11.45	3.23	1.30	3.36
2011-08	-6.73	-9.80	-9.20	-13.45
2011-12	-9.69	-0.40	-1.53	-4.72
2012-04	-0.78	-8.27	-6.24	-2.33
2012-08	11.25	6.25	3.10	1.34
2012-12	-0.49	4.48	1.43	2.94
2013-04	-10.07	-4.62	-6.53	-1.61
2013-08	1.56	-2.57	-1.79	0.29
2013-12	3.11	-1.09	-3.57	-0.96
Monthly average	1.55	0.88	0.83	0.34

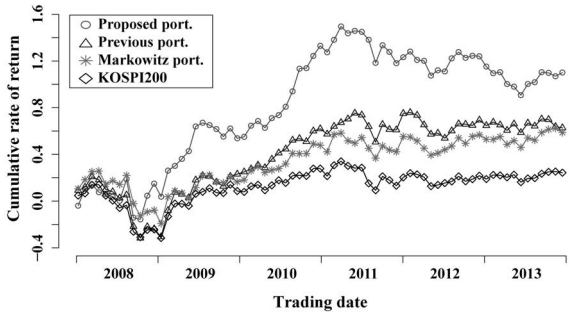


Fig. 9. Cumulative investment rate of return for each portfolio

모의 투자 실험을 위해 사용된 데이터는 2012년 12월 1일부터 2013년 11월 30일까지의 데이터이고, 모의 투자 기간은 2013년 1월부터 2013년 11월까지 총 11개월이다. 포트폴리오별 모의 투자 수익률은 Table 4에 정리되어 있다. 모의 투자 실험에서 실험 기간의 1개월 수익률 평균은 약 0.90%로, 기존 방법보다는 약 0.35%, 마코위츠 방법보다는 약 0.40%, 코스피 평균 수익률보다는 약 0.58% 높게 나타났다. '다음'과 '네이버' 주식토론방 데이터를 이용했을 때와 수익률은 다르지만 '네이버' 주식토론방 데이터를 이용한 경우에도 제안 방법으로 구성한 포트폴리오의 수익률이 비교 포트폴리오의 수익률보다 모두 높게 나타났다. 실험을 통해서 주식토론방 게시물과 주가와의 상관관계는 주식수익률과 어느 정도 관련이 있음을 알 수 있다.

특히, 이러한 특성은 특정 주식토론방에만 국한된 것이 아니라 인터넷상의 주식토론방에서 공통으로 나타나는 특성일 것으로 추측해볼 수 있다.

실험에서 사용한 포트폴리오들의 단순 수익률 비교만으로 제안 포트폴리오가 더 우수하다고 할 수는 없지만, 이 논문에서 주된 관심은 주식토론방에 나타난 사람들의 집단행위가 주가와 어떤 관련성이 있는지를 밝히고, 이러한 관련성이 주가 흐름을 예측하는 데 활용될 수 있음을 보이는 것이다. 주가와 주식게시판에 나타난 사람들의 집단행동 사이의 상관관계를 실제 주식투자에 활용하기 위해서 포트폴리오 구성 종목을 선택하기 위한 상관계수 임계치에 관한 통계적인 분석 및 추가적인 포트폴리오 위험성 분석이 뒷받침되어야 할 것이다.

6. 결 론

이 논문에서는 인터넷 주식토론방에 나타나는 사용자들의 집단행위와 주식시장 사이에 어떠한 상관관계가 있는지를 분석하였다. 이를 위하여 먼저 '다음' 포털의 주식토론방에 등록된 게시물을 수집하고 게시물 수의 변화와 주가와의 상관계수를 측정하였다. KOSPI200을 구성하는 주식 종목 중에서 57개 회사의 주식을 대상으로 분석한 결과, 주가 변화와 게시물 수 변화 사이에 뚜렷한 상관관계는 나타나지 않았다. 그렇지만 주가가 급락하거나 급등하는 시기에 게시물의 수가 증가하는 경향이 있으며, 주가와 게시물 수의 상관계수가 강한 양이거나 강한 음을 나타내는 경우 주식수익률도 그에 따라 높거나 낮은 경향이 나타났다.

이러한 분석 결과를 토대로 주식투자 종목 추천 알고리즘을 제안하였다. 제안 알고리즘은 기존 연구에서 제안하였던 주식 네트워크의 특성과 게시물과의 상관관계를 함께 고려하여 추천 종목을 선택한다. 마코위츠의 효율적 포트폴리오에 속한 종목들은 주식 네트워크에서 외곽에 위치하는 경향이 있다는 점을 이용하여 주식 네트워크에서 해당 종목의 차수가 낮고, 게시물과의 상관계수가 높은 종목을 추천 종목으로 선택하였다.

제안 방법으로 구성한 포트폴리오의 효율성을 보기 위해 57개 회사의 주식을 이용하여 모의 투자 실험을 수행하였다.

Table 4. The simulated investment earnings rate of our portfolio and three comparison portfolios (The proposed portfolio has been constructed using the article data of 'Naver.')

Period	Monthly average rate of return (%)			
	Proposed portfolio	Previous portfolio	Markowitz portfolio	KOSPI200
2013-01	-2.72	-4.83	-3.86	-2.85
2013-02	-2.65	1.21	-0.06	3.51
2013-03	-1.46	-1.77	2.28	-0.41
2013-04	1.27	-3.71	-6.53	-1.61
2013-05	-1.02	4.17	3.09	2.24
2013-06	1.35	-1.74	-5.74	-6.35
2013-07	11.01	10.57	8.05	3.14
2013-08	-0.42	0.38	-1.79	0.29
2013-09	1.77	2.59	6.41	3.75
2013-10	3.18	1.74	2.78	1.56
2013-11	-0.42	-2.53	0.87	0.27
Monthly average	0.90	0.55	0.50	0.32

포트폴리오 구성을 위해 '다음'과 '네이버' 포털에서 제공하는 주식토론방의 게시물 데이터를 이용하였다. 실험에서 '다음' 주식토론방 데이터를 이용하여 2008년 1월부터 2013년 12월까지의 데이터를 1개월 단위로 포트폴리오를 구성하여 실험한 결과, 제안 방법으로 구성한 포트폴리오의 1개월 평균 수익률은 약 1.55%로 주식 네트워크 특성만을 이용한 기존의 방법보다는 약 0.67%의 높은 수익률을 기록하였다. 또한, 마코위츠의 알고리즘으로 구성한 포트폴리오의 수익률과 KOSPI200 수익률보다 각각 약 0.72%와 1.21% 높게 나타났다. 또한, 모의 투자 기간 동안 제안 포트폴리오의 누적 투자수익률이 비교 포트폴리오들의 수익률보다 지속적으로 높게 나타났다.

'네이버' 주식토론방의 게시물 데이터를 이용한 모의 투자 실험에서는 2013년 1월부터 2013년 11월까지 1개월 간격으로 포트폴리오를 구성하였다. 제안 포트폴리오의 1개월 평균 수익률은 0.90%로 기존 방법으로 구성한 포트폴리오의 수익률보다 약 0.35% 높게 나타났으며, 마코위츠 포트폴리오와 KOSPI200의 수익률보다 각각 0.40%와 0.58% 높게 나타났다.

실제 주식투자를 위해서는 제안 방법으로 구성한 포트폴리오의 위험성 측정이나 주식토론방에 따른 가변적인 상관계수 임계치 설정 등에 관한 연구가 더 뒷받침되어야 하겠지만, 실험 결과를 통해 알 수 있는 것은 주식토론방에 게시물로 나타나는 사람들의 집단행위가 주가 흐름, 즉 주식시장의 변화를 어느 정도 반영하고 있으며, 이 데이터를 분석하여 주식투자에 활용할 수 있다는 것이다.

향후 연구로 포트폴리오 위험 분석과 실제 사용되고 있는 펀드 구성 종목과의 비교 등이 필요할 것이다. 또한, 이 논문에서는 게시물의 수에 대한 변화만을 고려하고 있으나 게시물 내용을 분석하여 주가 변화를 반영하는 단어나 감성이 있는지에 관한 연구가 추가된다면 인터넷상의 빅데이터를 활용하여 주식시장의 변화를 이해하는 데 도움이 될 것이다.

References

[1] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto, "Dynamics of market correlations: Taxonomy and portfolio analysis", *Physical Review E*, Vol.68, No.5, pp.1-12, 2003.

[2] G. Oh, C. Eom, F. Wang, W.-S. Jung, H. E. Stanley, and S. Kim, "Statistical properties of cross-correlation in the Korean stock market", *The European Physical Journal B*, Vol.79, No.1, pp.55-60, 2011.

[3] H.-J. Kim, I.-M. Kim, "Scale-free network in stock markets", *Journal of the Korean Physical Society*, Vol.40, No.6, pp.1105-1108, 2002.

[4] R. N. Mantegna, "Hierarchical structure in financial markets", *The European Physical Journal B-Condensed Matter and Complex Systems*, Vol.11, No.1, pp.193-197, 1999.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol.2, No.1, pp.1-8, 2011.

[6] M. Alanyali, H. S. Moat, and T. Preis, "Quantifying the relationship between financial news and the stock market", *Scientific Reports*, Vol.3, 2013.

[7] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using Google Trends", *Scientific Reports*, Vol.3, 2013.

[8] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis, "Quantifying Wikipedia usage patterns before stock market moves", *Scientific Reports*, Vol.3, 2013.

[9] H. Huh, S-H. Kim, S-K. Kang, and C-J. Eom, "Stock network an efficient portfolio in Korean stock market", *The Korea Journal of Financial Engineering*, Vol.5, No.2, pp.65-84, 2006.

[10] C-J. Eom, "An empirical study on the properties of stock network in the Korean stock market using the multifactor model and random matrix theory", *Journal of Industrial Economic Research*, Vol.20, No.5, pp.2055-2074, 2007.

[11] B. M. Tabak, T. R. Serra, and D. O. Cajueiro, "Topological properties of stock market networks: The case of Brazil", *Physica A: Statistical Mechanics and its Applications*, Vol.389, No.16, pp.3240-3249, 2010.

[12] W.-Q. Huang, X.-T. Zhuang, and S. Yao, "A network analysis of the Chinese stock market", *Physica A: Statistical Mechanics and its Applications*, Vol.388, No.14, pp.2956-2964, 2009.

[13] A. Vizgunov, B. Goldengorin, V. Kalyagin, A. Koldanov, P. Koldanov, and P.M. Pardalos, "Network approach for the Russian stock market", *Computational Management Science*, Vol.11, pp.45-55, 2014.

[14] T. Preis, D. Y. Kenett, H. E. Stanley, D. Helbing, and E. Ben Jacob, "Quantifying the behavior of stock correlations under market stress", *Scientific Reports*, Vol.2, 2012.

[15] Y.-J. Lee, G. Woo, "Analysis of the stock market network for portfolio recommendation", *Journal of The Korea Contents Association*, Vol.13, No.11, pp.48-58, 2013.

이 윤 정

e-mail : leeyj01@gmail.com

1995년 부경대학교 전자계산학과(학사)

1999년 부경대학교 전산정보학과(석사)

2008년 부경대학교 전자계산학과(이학박사)

2008년~현 재 부산대학교 BK21PLUS IT 기반융합산업창의인력양성사업단 박사후연구원

관심분야 : Computer graphics, Social network system analysis, Complex system

김 건 우

e-mail : gunwoo@pusan.ac.kr

2013년 부산대학교 정보컴퓨터공학부(학사)

2008년~현 재 부산대학교 전기컴퓨터공학과 석사과정

관심분야 : Static Analysis, Android, Go Language

우 균

e-mail : woogyun@pusan.ac.kr

2000년 한국과학기술원 전산학(박사)

2000년~2004년 동아대학교 컴퓨터공학과 조교수

2004년~현 재 부산대학교 전기컴퓨터공학과 교수

관심분야 : Programming language, Grid computing, Software metric, Program visualization, Complex system