빅데이터 트렌드를 이용한 주가지수 투자 전략

신 현 준*
라 현 우†

<초 록>

본 연구는 금융시장에 직간접적으로 참여하는 대중들의 심리(sentiment)의 변화와 시장 움직임 간에 연관관계가 있다고 전제하고 이를 주식시장 투자전략에 적용한다. 현대인의 삶은 인터넷과 연계되어있고, 크고 작은 금융위기는 시장 참여자의 인터넷을 통한 정보 검색에 영향을 미친다. 따라서 시장이 크게 움직이는 기간 동안 시장 참여자의 행동 – 인터넷 검색량에도 유의미한 변동이 있을 수 있다고 제안하고, 빅데이터에 내재된 트렌드를 분석하여 이를 검증한다. 경제 및 금융과 관련된 용어들에 대해 네이버 트렌드 툴로써 검색량 변화 패턴을 찾아내고, 이것을 주식시장 움직임의 선제적 신호로 판단하여 투자전략을 구성한다. 연구결과 빅데이터 트렌드 분석은 경제활동에 참가하는 사람들의 집단행동을 이해하고 이를 투자에 효과적으로 활용 가능함을 보인다.

* 교신저자, 상명대학교 경영공학과, hjshin@smu.ac.kr
† 상명대학교 경영공학과, hwra@smu.ac.kr

Ⅰ. 서 론

대중들의 일상생활에 내재된 행동양식을 반영하는 빅데이터의 증가는 과학자들에게 복잡한 세계에 대한 근본적인 문제를 해결하기 위한 새로운 도전 기회가 되고 있으며, 금융시장 역시 이러한 과학적 접근방법의 주요 도전 대상중 하나이다. 예상치 못한 시장의 움직임은 개인의 재산 및 경제적 사안들에 큰 영향을 미친다. 따라서 금융시장을 새로운 시각에서 모델링하려는 시도와 함께 네트워크 콘텐츠 분석을 통해 시장 움직임을 포착하기 위한 연구가 최근 활발히 진행되고 있다. 실제로 금융 거래 데이터에는 시장 참여자의 수많은 결정 사항들이 반영되어있으며, 과거 글로벌 금융 위기 시의 매매 의사결정 및 거래패턴과 관련한 시장데이터는 이러한 위기들을 초래한 복잡한 인간행동의 단면을 보여준다.

일반적인 의사 결정과정은 사전 정보 수집에서 시작되고, 주로 온라인 인터넷 검색을 통해 필요한 정보들이 수집된다. 최근 글로벌 검색 엔진인 구글(Google)은 구글 트렌드(Google Trends)[1]라는 툴을 통해, 서로 다른 검색어들에 대한 질의(query)의 양(volume), 다시 말해 검색량(search volume)과 시간에 따른 검색량 변동 데이터에 대한 서비스를 제공하기 시작했다. 마찬가지로 국내에서도 검색 엔진 국내시장점유율 1위인 네이버 (Naver)에서 네이버 트렌드[2] 서비스를 제공하고 있다.

구글 트렌드 데이터를 분석한 최근 연구들은 특정 용어들의 검색량 변화가 인플루엔자 감염 환자 수 증가 추세 또는 주식시장의 거래량 변화 등과 관련이 있다는 것을 보인 바 있다(Goel et al. 2010). 주식시장 거래량과 검색량 간의 연관성은 야후(Yahoo) 데이터를 통해서도 입증이 되었으며, Choi and Varian(2012)은 구글 트렌드 데이터가 자동차 판매량, 실업수당 신청률, 소비자신뢰지수 등 여러 경제지표와 연계될 수 있다고 주장하였다. 투자자는 시장불안감이 증폭되는 기간 동안에는 주식 매입 또는 매도결정이전에 시장에 대한 정보 검색을 평소보다 더 많이 실시한다는 Preis et al.(2013)의 주장이 이를 뒷받침한다. 그러나 구글이 미국뿐만 아니라 세계 시장 점유율 1위의 검색서비스를 제공하고 있는 반면에, 국내 시장 점유율은 5%미만에 머무르고 있어 국내 주식시장과 구글 트렌드 검색량 간의 연계성을 검증하는데 한계가 있다.[3]

시장 참여자가 경제에 갖는 관심은 인터넷 사용자의 경제 관련 검색어와 연결될 수 있으며 특정 기간의 해당 검색량은 경제 분야 빅데이터의 트렌드로 이해할 수 있다. 따라서 본 연구는 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 가설을 수립한다. 이러한 가설 하에 국내 주식시장의 주가 및 네이버 트렌드(Naver trends) 검색량 데이터 변화량의 연관성을 이용한 세 개의 주가지수 투자전략을 제안하고 가설을 검증하고자 한다. 또한 사용자가 정적환경에서 생성하는

1) http://www.google.co.kr/trends/
2) http://trend.naver.com/
3) 출처: http://www.internettrend.co.kr/ (점유율 산출 기간: 2010.1.1. ~ 2014.7.31.)

빅데이터 트렌드는 PC 기반의 검색량으로, 동적환경에서 생성하는 트렌드는 모바일(mobile) 기반의 검색량으로 각각 구분하여 주식시장의 움직임에 미치는 영향을 비교한다. 스마트폰 등의 모바일기기가 빅데이터가 이슈화되고 성장하는데 크게 기여해 왔음은 주지의 사실이므로, 모바일 기기를 통해 생성된 빅데이터 트렌드가 주식시장 참여자의 주요 의사결정과도 관련이 있는지 여부를 관찰하는 것도 의미 있을 것이라 기대한다. 이와 더불어 특정 용어의 검색량 시계열이 주가지수 시계열에 대해서 갖는 선행성 유무를 Granger인과관계(Granger casuality) 분석을 통해 파악하고 인과관계의 정도가 투자성과에 미치는 영향에 대해서도 분석하고자 한다.

Ⅱ. 관련문헌연구

개인 및 기관투자자의 투자심리를 투자전략에 활용한 연구는 크게 두 분야로 정리할 수 있다. 하나는 실제 금융거래 데이터에 드러난 심리적인 요인들을 분석하여 전략을 수립하는 직접적인 방식이고 다른 하나는 본 연구와 같이 빅데이터 등의 트렌드를 분석하여 투자 전략을 수립하는 간접적인 접근방식이다.

먼저, 직접적인 심리투자 전략에 대한 연구로써 Kumar and Lee(2006)는 미국 주식시장에서 개인투자자의 집단적 거래행태를 발견하고, 개인투자자의 투자심리(investor sentiment)와 주식수익률 간의 유의한 관계가 있음을 보였다. Jackson(2003)과 Dorn et al.(2008)은 호주와 독일 주식시장에서 브로커를 통한 개인투자자들의 거래를 각각 분석함으로써 개인투자자들 간의 거래에 유의한 상관관계가 있음을 밝혔다. 개인투자자들은 매수나 매도의 한 방향으로 집중적인 거래를 하는 특징을 보이는데, Barber et al.(2009)는 이러한 특징을 체계적 잡음 (systematic noise)라 정의하고, 개인투자자들의 심리적 요인(psychological bias)에 의해 나타난다고 주장하였다. Kang et al.(2013)은 국내주식시장에서 개인투자자의 거래비중이 높은 주식일수록 개인투자심리에 의한 가격 변화 이후 점차적인 가격 반전(price reversal)이 관찰됨을 보고하였다.

빅데이터나 매체로부터 대중의 심리를 추론하여 간접적인 투자전략을 수립하는 최근 연구에는 Bollen et al.(2011), Preis et al.(2010), Bordino et al.(2012), Goel et al.(2010) 등이 있다. Bollen et al.(2011)은 감정이 개개인의 행동과 의사결정에 큰 영향을 미친다는 행동경제학 (behavioral economics) 이론을 토대로 대규모 트위터 피드 (twitter feeds)로부터 추출한 집단적 감정상태의 변화가 일정기간동안 다우존스산업지수 (DJIA)의 가치 변화와 상관관계가 있음을 밝혔다. Preis et al.(2010)은 2004년부터 2010년까지 구글 검색엔진을 통해 질의된 다양한 검색어들의 주별 (weekly) 검색량을 토대로 이들 검색량과 주식시장의 움직임간의 상관관계가 있음을 보고하였다. 특히 S&P500에 포함된 개별 종목들의 검색량이 개별주식의 거래량과도 유의한 상관관계가 있다는 것을 보였다. Bordino et al.(2012)도 2010년 5월부터 2011년 4월까지 야후 (Yahoo) 검색엔진을 통해 질의된 검색어의 검색량과 NADAQ -100에 포함된 종목들의 거래량 간의 상관관계를 보였고, 이들 두 시계열

간의 인과관계를 분석하였다. Goel et al.(2010)은 웹검색 분석을 통해 소비자 행동을 예측할 수 있음을 빌보드 순위, 독감 유행 예측, 음반 및 영화표 판매 등을 통해 입증하였다.

빅데이터와 관련한 국내의 선행 연구로써 Kim et al.(2012)은 뉴스 컨텐츠를 분석하기 위해 오피니언 마이닝이라는 빅데이터 감성분석 기법을 통한 지능형 투자의사결정모형을 제시하여 뉴스 컨텐츠의 감성분석 결과와 주가지수 등락이 유의한 관계를 가지고 있음을 보였고, 분석결과를 이용해 주가지수 변동성 예측이 가능할 것으로 판단하였다. Ahn et al.(2010)은 남북관계와 관련되는 뉴스와 주식시장의 상관관계 분석을 통해 뉴스가 주가지수 가격변동에 유의미한 영향을 미친 것으로 분석하였다. Park(2012)은 과학적 데이터, 즉 통계나 실험을 통해서 얻은 정형화된 데이터뿐만 아니라 인간의 정서나 심리 정보에 해당하는 기분이나 감정이 내재된 비정형화된 데이터를 분석함으로써 소비자 중심의 정보를 산출할 수 있으며, 기업은 물론 공공 영역에서도 광범위하게 사용될 수 있다고 판단하였다. Lee et al.(2014)은 빅데이터에 나타난 투자자별 감성이나 정보가 KOSPI200 선물 지수 수익률에 미치는 영향을 실증 분석하여 빅데이터가 KOSPI200 선물 지수 수익률을 예측하는 정보를 포함한다는 것과 빅데이터를 사용한 투자 전략이 높은 수익을 가져옴을 증명하였다.

이상의 기존연구를 고찰해보면 우리나라 인터넷 포털 검색엔진에서 형성되는 빅데이터 트렌드를 이용한 국내 주식시장 투자전략에 대한 연구는 찾아보기 어렵다. 따라서 본 연구는 국내 대표적인 검색엔진을 통해 형성되는 사용자의 경제관심의 변화 트렌드를 주가지수에 반영하는 투자전략을 제안하고 그 성과를 분석하고자 한다. 본 논문의 3장에서는 경제관련 검색어의 검색량 변화 트렌드를 이용한 세 개의 KOSPI 주가지수 투자전략에 대해서 설명한다. 4장에서는 제안한 세 개의 전략들을 PC와 모바일 환경으로 구분하여 분석하고, 랜덤투자전략 및 매수유지전략과의 비교를 통해 투자성능을 검증한다. 또한 검색량 시계열이 KOSPI 지수 시계열에 대해 선행성을 갖는지 여부를 분석하고 선행성이 투자성과에 미치는 영향을 파악한다. 마지막으로 5장에서는 본 연구의 결론을 정리한다.

Ⅲ. 연구의 자료 및 투자전략

본 연구는 정적인 기기인 ‘데스크탑 및 노트북 PC에서의 검색’과 ‘모바일 기기에서의 검색’의 두 경우로 구분하여 투자자의 주식시장 관심도가 주가에 미치는 영향을 분석한다. 자료가 제공되기 시작한 각각 2007년 1월과 2010년 7월부터 2014년 3월까지의 기간 동안 네이버 트렌드 검색량을 수집하였으며 정적인 PC와 동적인 모바일 기기를 통한 검색이 투자전략 성과에 미치는 차이점을 실험을 통해 분석하도록 한다. 투자전략의 대상이 되는 검색 용어는 임의로 선택하지 않고 금융 및 경제와 관련된 용어들로 국한하여 선정하였다. 전체 용어의 일부는 관련용어를 생성해주는 구글의 세트(sets)를 이용하고 나머지 용어들은 네이버의 연관검색어를 이용하여 총 92개의 검색어들로 구성한다. 투자전략은 실제 시장을 추종하는 ETF상품을 활용할 수 있으나, 본 연구에서는 분석의

단순화를 위해 코스피(KOSPI) 종합주가지수를 대상으로 하였으며 공매도(short selling)가 가능하다고 가정한다.

특정 용어에 대한 검색량과 국내 주식 투자자들의 매매 방향성 간의 관련성을 파악하기 위해, 코스피 지수의 t주의 첫 거래일 종가인 $p(t)$를 분석한다. 예를 들어 $t-1$주에 CMA 용어가 네이버에서 검색된 검색량, $n(t-1)$을 네이버 트렌드를 통해 산출할 수 있다. 투자자들의 정보검색 움직임을 정량화하기 위해 Preis et al.(2013)의 연구에서 사용한 검색 상대변화량(relative search volume change) $\Delta n(t,\Delta t) = n(t) - N(t-1,\Delta t)$ 을 이용한다. 여기서 $N(t-1,\Delta t)$ 는 $\{n(t-1)+n(t-2)+...+n(t-\Delta t)\} \ /\Delta t$로 정의되며 $t-1$주부터 최근 Δt주 동안의 검색량의 이동평균을 뜻한다. <그림 1>은 KOSPI 지수 종가와 검색용어 CMA에 대한 검색 상대변화량의 시간에 따른 변화를 보여주며, 각각은 KOSPI 지수 t주 첫 거래일의 종가, 즉 $p(t)$의 시계열과 검색용어 CMA의 Δt = 3주로 산출한 검색 상대변화량을 뜻한다. 2007년 1월 5일부터 2014년 3월 28일까지의 기간 동안 이러한 직접/간접 투자자들의 정보검색 행동 패턴의 변화가 주가 변동으로 이어진다는 가설을 검증하기 위해서, 본 연구에서는 3개의 투자전략을 제안한다.

3개의 전략은 공통적으로 시장의 미래 방향성 예측이 정확해야만 수익을 낼 수 있는 구조이고, 특히 변동성이 큰 시장에서 더 많은 수익을 내는 특성을 갖는다. 또한 3개의 전략 모두 동일한 매매신호를 이용하지만, 청산방법에 따라 일시청산 (en bloc settlement; 이하 EBS), 일대일청산 (one-to- one settlement; 이하 OS) 그리고 독립청산 (independent settlement; 이하 IS)로 구분되며 네이버 트렌드를 이용한다는 의미에서 세 투자전략을 각각 NT-EBS전략, NT-OS전략, NT-IS전략으로 정의하고 각 전략에 대한 설명은 다음과 같다.

1. NT-EBS전략

매도와 매수신호는 주 단위(weekly)로 각각 $\Delta n(t-1,\Delta t)$ 가 0보다 클 때와 작을 때 발생하고, 매매는 t주의 첫 거래일 종가인 $p(t)$를 기준으로 한다. NT-EBS전략은 매매신호 발생 시에 취한 포지션을 바로바로 청산하지 않고 보유한다. 청산은 기존의 유지하고 있는 포지션과 반대의 매매신호가 발생하면 한꺼번에 시행한다. 예를 들어, 만약 $t-1$주에 $\Delta n(t-1,\Delta t) > 0$ 이라면 t주의 첫 거래일 종가인 $p(t)$에 매도하고 만약, $t+1$주에 $\Delta n(t+1,\Delta t) > 0$ 이라면 $t+2$주의 첫 거래일 종가인 $p(t+2)$에 다시 한 번 매도포지션을 취한다. 그리고 만약 $t+2$주에 $\Delta n(t+2,\Delta t) < 0$ 이라면 $t+3$주의 첫 거래일 종가인 $p(t+3)$로 2회 매수함으로써 기존에 누적된 2개의 매도 포지션들을 동시에 청산한다. 본 연구에서는 로그수익률을 이용하여 누적수익률을 산출하므로 위 예의 경우 누적수익률은 $\{\log p(t) - \log p(t+3)\} + \{\log p(t+2) - \log p(t+3)\}$ 이 된다.

2. NT-OS전략

NT-OS전략은 매매신호 발생 시에 취한 포지션을 바로 청산하지 않고 보유해 나가는 방식은

NT-EBS전략과 동일하지만, 청산은 기존의 유지하고 있는 포지션과 반대의 매매신호가 발생하면 보유 중인 가장 오래된 반대 포지션과 일대일로 시행한다. 마찬가지로 예를 들어, 만약 $t-1$ 주에 $\Delta n(t-1,\Delta t) > 0$ 이라면 t주의 첫 거래일 종가인 $p(t)$에 매도하고 만약, $t+1$ 주에 $\Delta n(t+1,\Delta t) > 0$ 이라면 $t+2$ 주의 첫 거래일 종가인 $p(t+2)$에 다시 한 번 매도포지션을 취한다. 그리고 만약 $t+2$ 주에 $\Delta n(t+2,\Delta t) < 0$ 이라면 $t+3$ 주의 첫 거래일 종가인 $p(t+3)$로 매수함으로써 기존에 누적된 매도 포지션들 중 t주의 매도포지션과 일대일로 청산된다. 이 경우 누적수익률은 $\log p(t) - \log p(t+3)$ 이다.

3. NT-IS전략

NT-IS전략은 매매신호 발생 시에 취한 포지션을 일주일만 보유하고 차주 첫 거래일에 청산한다. 예를 들어, 만약 $t-1$ 주에 $\Delta n(t-1,\Delta t) > 0$ 이라면 t주의 첫 거래일 종가인 $p(t)$에 매도하고 $t+1$ 주의 첫 거래일 종가 $p(t+1)$로 매수하여 청산한다. 반대로 만약 $t-1$ 주에 $\Delta n(t-1,\Delta t) < 0$ 이라면 t주의 첫 거래일 종가인 $p(t)$에 매수하고 $t+1$ 주의 첫 거래일 종가 $p(t+1)$로 매도하여 청산한다. 이 경우 누적수익률은 각각 $\log p(t)$ -$\log p(t+1)$ 과 $\log p(t+1)$ -$\log p(t)$ 이다. 제안한 3개의 전략 중 거래가 빈도가 가장 많은 IS전략의 경우에도 일 년에 최대 104회 미만의 거래가 발생 가능하므로 투자전략을 사용하는 과정에서 거래비용은 무시한다.

NT-EBS전략과 NT-OS전략은 일종의 레버리지 전략으로써 동일한 방향의 매매 신호가 발생하면 동일한 포지션을 누적해가기 때문에 방향성 예측이 정확하다면 큰 수익을 얻을 수 있지만 반대의 경우 큰 손실의 위험도 존재한다. 레버리지의 강도를 비교하면 기존의 보유 포지션과 반대 신호 발생 시 전량 청산하는 NT-EBS전략과 달리 NT-OS전략은 가장 먼저 취한 포지션만 일대일로 청산하기 때문에 더 큰 레버리지를 갖는다고 할 수 있다. 반면 일주일단위로 포지션을 청산하는 NT-IS전략은 3개의 전략 중 가장 작은 레버리지를 갖는다.

Ⅳ. 실증분석 결과

1. 전략별 투자성과 분석

검색 상대변화량의 $\triangle t$를 3으로 설정하고 네이버 트렌드의 PC 검색용어들 중에서 'CMA'를 예로 들어서, <그림 2>는 본 연구에서 제안한 3개 전략과 매수유지(buy and hold)전략 및 랜덤투자(random investment)전략들 간의 누적수익률 손익을 각각 비교하고 있다. 매수유지전략은 투자대상기간이 시작되는 2007년 1월 5일 KOSPI200 지수를 종가로 매입한 후 포지션을 계속 유지하는 전략이다. 랜덤투자전략은 NT-EBS, NT-OS, NT-IS전략과 마찬가지로 매매신호가 발생되는 시점에서 매매하지만, 매매신호의 방향과는 상관없이 랜덤으로 매수와 매도를 결정하는 전략이며, 청산하는

방법에 따라 각각 Rand-EBS, Rand-OS, Rand-IS전략으로 구분된다.

Rand-EBS, Rand-OS, Rand-IS 전략별로 10,000회 시뮬레이션을 수행하여 얻은 누적수익률 평균은 각각 -1.97%, -0.11%, -1.86% 이고 표준편차는 각각 122.73%, 370.0%, 65.45%이었다. 레버리지가 가장 큰 전략인 NT-OS와 동일한 청산방법을 갖는 Rand-OS 역시 매우 큰 표준편차를 나타냄을 알 수 있다. 누적수익률은 NT-IS, NT-EBS, NT-OS전략 순으로 각각 113%, 237.5%, 1019.5%, 그리고 매수유지전략은 35.0%의 결과를 보였다. <그림 2>의 차트 (A), (B), (C)는 차례대로 NT-EBS, NT-OS, NT-IS전략과 각각의 랜덤투자전략 및 매수유지(buy and hold) 전략들 간의 누적수익률 손익결과를 보여준다. 본 예시의 검색용어는 CMA이고, Δt=3이다. 매수유지전략은 세 차트에서 35.0%의 동일한 누적수익률을 나타낸다. 청산방법에 따라 랜덤투자전략도 Rand- EBS, Rand-OS, Rand-IS로 구분하며 괄호안의 Mean, +Std, -Std는 각각 10,000회 랜덤실험에서의 평균, 평균+표준편차, 평균-표준편차를 의미하고, 세 랜덤전략 모두 평균은 0에 근접함을 알 수 있다.

NT-EBS, NT-OS, NT-IS의 세 전략별로 92개의 전체 검색어에 대한 투자성과를 각각 정적환경(PC)과 동적환경(모바일기기)으로 구분해 수익률 순으로 정렬한 결과는 <그림 3-4>와 같다. 여기서 검색 상대변화량 계산 시 이동평균기간으로 사용되는 Δt는 2주로 설정하였다. <그림 2>와 같이 전략별로 매우 큰 편차를 보이는 누적로그수익률의 상대비교를 위해 본 연구에서는 각 전략의 랜덤투자(Rand-EBS, Rand-OS, Rand-IS)를 중심으로 정규화 누적로그수익률을 사용하도록 한다. 즉, PC와 모바일 별로 각 전략의 수익률에서 10,000회 시뮬레이션 한 랜덤투자의 수익률 평균을 빼고, 이를 랜덤투자의 수익률 표준편차로 나눠서 산출한다(예, {검색어 'CMA'의 NT-EBS 수익률 - Rand-EBS 수익률 평균}/Rand-EBS 수익률 표준편차). 그리고 앞서 기술하였듯이 10,000회 시행한 랜덤투자전략의 평균은 0에 근접하였으므로 <그림 3>의 (A)와 같이 중앙에 0(랜덤투자 평균)을 중심으로 좌우로 -2σ, -σ, σ, 2σ 등의 점선으로 구분하고 있다. 매수유지전략과도 비교하기 위해 매수유지전략의 정규화 누적로그수익률도 각 차트에 포함돼 있으며, 모바일 검색어 기반 전략보다는 PC 검색어 기반 전략들에서 확연히 하단에 위치해 있는 것을 확인할 수 있다.

본 연구에서 제안한 세 전략들의 결정모수에 대한 강건성(robustness)을 실험하기 위해 정규화 누적로그수익률들을 모든 이동평균기간(Δt=1,...,6)에 대해 각각 산출하여 그 결과를 <표 1>에 정리하였다. 이 표는 제안한 전략들의 이동평균기간에 따른 정규화 누적수익률을 PC와 모바일로 구분하여 나타내고 있다. 각 이동평균기간의 숫자들은 전체 92개 검색어들의 정규화 누적로그수익률 평균을 의미하고, ALL 열은 전체 이동평균기간(Δt=1~6)의 평균이다. 랜덤투자전략과의 비교를 위해 단일표본 t-검정(귀무가설: 제안전략 수익률 평균 = 랜덤전략 수익률 평균, 대립가설: 제안전략 수익률 평균 > 랜덤전략 수익률 평균)을 시행하였다. 괄호안의 숫자는 t값이고 *, **, ***는 각각 p값 < 0.05, 0.01, 0.001일 경우에 해당한다.

PC의 경우 전반적으로 세 전략모두 랜덤투자전략보다 매우 유의하게(p값 < 0.001) 높은 수익률을 보이고, 모바일의 경우 NT-OS전략만 유의하게 높은 수익률을 나타냄을 관찰할 수 있다. PC 검색어를

기반으로 한 실험결과가 전반적으로 모든 전략들에서 랜덤투자전략보다 매우 유의하게 높은 수익률을 보임을 관찰할 수 있다(NT-EBS: 정규화 수익률 평균 = 0.29, t = 5.52, df = 551(= 6×92 - 1), p < 0.001; NT-OS: 수익률 평균 = 0.93, t = 23.4, df = 551, p < 0.001; NT-IS: 수익률 평균= 0.30, t = 9.29, df = 551, p < 0.001, ALL의 경우임). NT-EBS 및 NT-IS 전략은 Δt=2일 때 가장 우수한 성과를 보였다. 모바일 검색어를 기반으로 한 결과에서는 NT-OS 전략의 ALL의 경우(수익률 평균 = 0.07, t = 2.01, df = 551, p < 0.05)를 제외하면 전반적으로 세 전략 모두 랜덤투자전략 대비 유의한 성능을 보이지 않는 것으로 나타났다. 이와 같은 결과는 서론에서 언급한 PC로부터 생성되는 빅데이터 트렌드가 모바일 기기로부터 생성되는 빅데이터 트렌드와는 달리 주식시장에 매우 유의한 영향을 미친다는 것을 확인할 수 있다. 모바일 기기로부터 생성되는 빅데이터 트렌드가 PC의 경우보다 주식시장의 주가 움직임에 직접적인 영향을 미치지 못하는 것은 다음의 원인들로 추측해 볼 수 있다. 하나는 사용자가 모바일 기기를 사용하는 주된 목적이 사진 및 동영상 촬영 및 공유, 소셜네트워킹, 음악 청취 또는 단순 검색 등의 여가활동이라는 점이고, 다른 하나는 PC로부터의 인터넷 트래픽 발생시간이 주로 오전8시부터 오후4시에 치중되어있는 반면, 모바일 기기 중 스마트폰 및 태블릿으로부터의 트래픽 발생시간이 각각 자정부터 오전8시까지와 오후4시부터 자정까지에 집중되어 금융시장이 열리는 시간과 분리돼 있다는 것이다.[4)]

본 연구에서 제안한 세 전략 모두 청산 방법에 차이가 있을 뿐 기본적으로 경제관련 특정 검색어의 검색량이 증가하면 주가지수를 매도하고, 감소하면 매입한다는 것을 전제로 한다. 제안 전략들이 유의한 수익률을 산출한다는 결과는 검색량 증가 후 시장 주가가 하락하고, 감소 후 시장주가가 상승할 것이라는 가설을 본 실험의 실증 결과로 입증하는 것이다. 다시 말하면, 시장 참여자가 경제에 갖는 관심은 인터넷 사용자의 경제 관련 검색어와 연결될 수 있으며 특정 기간의 해당 검색량은 경제 분야 빅데이터의 트렌드로 이해할 수 있다. 따라서 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다고 할 수 있으며, 이와 같은 국내 빅데이터 트렌드와 주식시장 간의 관계는 Preis et al.(2013)의 미국 주식시장의 결과와도 일치한다.

Ilaria et al.(2012)와 Preis et al.(2010) 및 Goel et al.(2010) 등은 웹 검색량이 주식시장 거래량에 영향을 미치고 이를 예측할 수 있다고 주장한 바 있다. 본 연구에서 PC 검색어와 모바일 검색어의 검색량을 기준으로 코스피시장의 거래량간의 상관관계를 각각 분석한 결과 PC 검색량과 코스피시장의 거래량은 유의한 양의 상관관계(Kendall's tau = 0.093, z = 2.71, N = 92, p < 0.005)를 보였으나 모바일 검색량과 코스피시장의 거래량 간에는 유의한 관계를 보이지 않았다. 이것은 위 실험결과에서 관찰한 PC 검색어 기반의 전략들과 모바일 검색어 기반의 전략들 간의 수익률 결과와도 유관하다. 특히 Ilaria et al.(2012)은 웹 검색량의 변화가 주식시장 거래량 증감에 미치는

4) 출처: the smart data manifesto (2013.4.30.) 보고서의 4페이지 및 6페이지, http://exelate.com/insights/uncategorized/smart-data-series-smart-data-manifesto/

기간이 일주일 이내 즉, 1~2일 이내라고 제시하였는데, 본 연구의 실험(PC 기반)에서도 주단위로 전체 검색어의 검색량 변화가 코스피시장의 거래량 증감에 미치는 기간은 해당 주 즉, 1주 이내라는 유사한 결과를 얻었다.

2. 검색량과 주가지수 시계열간의 선행성에 따른 투자성과 분석

본 연구에서는 결과가 원인에 앞설 수 없다는 원인의 시간선행성에 기반을 둔 Granger인과관계 분석을 통해, 92개의 검색어 검색량 시계열들과 KOSPI 지수 간의 선행성을 검증하고 선행성이 투자성과에 어떤 영향을 미치는지 파악한다. Granger인과관계에 대한 설명은 송유철, 원용걸(2011) 및 김용순 외(2011) 등의 기존 연구에 자세히 기술되어 있으므로, 검증 절차는 간단히 설명하도록 한다. 본 연구는 Toda and Yamamoto(1995)의 부가된 벡터자기회귀(augmented VAR) 모형을 사용하며, 이를 TY모형이라 명명한다. TY모형은 단위근 검정이나 공적분 검정 등 일관성이 부족한 사전 검정 결과와 관계없이 Granger인과관계 검정을 실시할 수 있어 시계열의 안정성에 대해 어느 정도 강건한 인과성 검정방법이다.

먼저 시계열들이 안정적(stationary)인지 여부를 판단하여 차분계수(order of integration)를 결정하기 위해서, 특정 검색어의 검색량 시계열과 KOSPI 지수 시계열을 대상으로 ADF(augmented Dickey-Fuller)검정을 실시한다. 여기서 결정된 최대 차분계수를 $m(m$=0,1,2)이라 한다. 그 다음 VAR모형을 수립하고 최적시차(optimal lag) p를 결정한다. VAR모형의 최적 시차, p를 결정하기 위해 AIC(아카이케 정보기준), SIC(슈와르츠 베이지안 정보기준), HQ(하난-퀸 정보기준)의 정보기준이 일반적으로 가장 많이 사용되며, 본 연구는 안전한 결과를 채택한다는 의미에서 세 정보기준 중 가장 긴 최적시차를 산출하는 AIC 정보기준을 사용하였다. Granger인과관계 검정은 벡터자기회귀 모형인 VAR($m+p$)을 이용하여 최적시차인 p개 계수에 대해서만 카이제곱(χ^2) 검정을 실시하는 수정된 왈드검정(modified Wald test)을 통해 수행된다.[5)]

수행결과 전체 92개의 검색어의 검색량 시계열들 중 28개, 즉 30.4%의 시계열이 1%의 유의수준으로 KOSPI 지수 시계열을 Granger초래하는 것으로 나타났고, 22.8%의 시계열과 14.1%의 시계열이 각각 5%와 1%의 유의수준으로 KOSPI 지수 시계열을 Granger초래하였다. TY모형의 수정된 왈드검정은 카이제곱 통계치(χ^2 statistics) 및 p값(p-value) 등을 산출해 주는데, 시계열의 선행성과 투자성과 간의 상관관계를 분석하기 위해 p값을 이용한다. <표 2>의 Panel A는 52개 검색어들의 검색량 시계열과 KOSPI 주가지수 시계열 간의 Granger인과관계 검증(귀무가설: 특정 검색어의 검색량 시계열이 KOSPI 주가지수 시계열을 Granger초래하지 않는다)을 통해 얻은 p값과 PC기반의 NT-EBS, NT-OS, NT-IS 전략들의 수익률 간의 Kendall rank correlation 검증 결과를 보여준다. Kendall 검증 결과에 의하면 상관도를 의미하는 tau값이 모든 전략에 대해 통계적으로

5) Granger인과관계 검정을 위한 최적시차 결정 및 변형된 왈드검정은 R에서 제공하는 패키지를 사용하였고, 구현은 다음 사례를 참고하였다. http://www.christophpfeiffer.org/2012/11/07/toda-yamamoto-implementation-in-r/

유의한 음의 상관관계를 나타낸다. 이것은 검색량 시계열이 KOSPI 주가지수 시계열을 Granger초래할수록, 즉 선행성이 있을수록 높은 수익률을 산출한다는 것을 의미한다. Panel B는 Granger인과관계 검증에서 수정된 왈드검정 결과인 p값의 구간별 평균 수익률을 보여주며, 각 구간의 대푯값을 각각 0.005, 0.025, 0.075, 0.3으로 설정했을 때의 Pearson 상관계수 값을 전략별로 나타낸다. 이 상관계수가 −1에 가까운 값을 갖는다는 것은 검색량 시계열이 KOSPI 주가지수 시계열에 대한 선행성이 높을수록 (수정된 왈드검정의 p값이 작을수록) 높은 수익률을 산출한다는 것을 뜻하므로, Panel A의 결과에 부합함을 알 수 있다.

Ⅴ. 결 론

본 연구는 대중들의 감정이 개인행동과 의사결정에 큰 영향을 미칠 수 있다는 행동경제학의 이론을 토대로 국내 빅데이터 트렌드를 이용한 KOSPI 주가지수 투자전략을 제안하였다. 시장 참여자가 경제에 갖는 관심은 인터넷 사용자의 경제 관련 검색어와 연결될 수 있으며 특정 기간의 해당 검색량은 경제 분야 빅데이터의 트렌드로 이해할 수 있다. 따라서 본 연구는 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 가설을 수립하였고, 국내 주식시장의 주가 및 네이버 트렌드(Naver trends) 검색량 데이터 변화량의 연관성을 이용한 세 개의 주가지수 투자전략을 제안하고 가설을 검증하였다. 또한 사용자가 정적환경에서 생성하는 빅데이터 트렌드는 PC 기반의 검색량으로, 동적환경에서 생성하는 트렌드는 모바일(mobile) 기반의 검색량으로 각각 구분하여 주식시장의 움직임에 미치는 영향을 비교하였다. 실험결과 제안한 전략들(NT-EBS, NT-OS, NT-IS) 모두 랜덤투자 및 매수유지 전략보다 매우 유의한 투자 수익률을 보였고, 세 전략 중에는 NT-OS 전략이 가장 우수한 결과를 나타내었다. 스마트폰 등의 모바일 기기가 빅데이터를 이슈화시키고 성장하게 한 것과는 다르게, 모바일 기기로 부터의 생성된 경제분야 빅데이터 트렌드는 주식시장 참여자의 주요 의사결정에는 큰 영향을 주지 않는 것으로 분석됐다. 마지막으로 92개 검색어들의 검색량 시계열이 KOSPI 주가지수에 대한 선행성이 있을수록 더 좋은 투자성과를 보이는 것으로 Granger인과관계 검증을 통해 파악되었다. 이것은 본 연구에서 수립한 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 가설의 타당성을 뒷받침한다. 향후 검색량 시계열과 주가 시계열 간의 선행성 분석을 통해 최적 시차(optimal lab)에 기반을 둔 거래 전략의 개발 및 KOSPI 지수 투자 이외에 빅데이터 트렌드와 상장지수펀드(ETF) 간의 결합을 통한 전략 개발에 대한 추후 연구가 계획되어 있다.

참 고 문 헌

송유철, 원용걸, “동아시아 국가들의 실질환율, 순수출 및 경제성장간의 상호관계 비교연구: 시계열 및 패널자료 인과관계 분석”, 「KIF Working Paper」, 2011-08, 1-47.

김용순, 권치홍, 이경애, 이현림, “2008년 금융위기 이후 부동산가격 결정요인 변화 분석”, 「LHI Journal」, 2, 4, 2011, 367-377.

Ahn, H. J., Jeon, S. P. and Chay, J. B. (2010), The Effects of News Related to the North-South Korean Relationship on the Korean Stock Markets, KIF Analysis on Korean Economy, 16(2), 199-231.

Barber, B. M., Odean, T. and Zhu, N. (2009), Systematic Noise, Journal of Financial Markets, 12, 547-569.

Bollena, J., Maoa, H. and Zeng, X. (2011), Twitter mood predicts the stock market, Journal of Computational Science, 2(1), 1-8.

Bordino, I., Battiston, S., Caldarelli, G., Cristell, M. and Ukkonen, A. (2012), Web Search Queries Can Predict Stock Market Volumes, PloS ONE, 7(7), 1-17.

Choi, H. and Varian, H. (2012), Predicting the Present with Google Trends. The Economic Record, 88, 2-9.

Dorn, D., Huberman, G. and Sengmueller, P. (2008), Correlated Trading and Returns, Journal of Finance, 63, 885-920.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. and Watts, D. J. (2010), Predicting consumer behavior with Web search, PNAS, 107(41), 17486-17490.

Jackson, A. (2003), The Aggregate Behaviour of Individual Investors, Available at SSRN: http://dx.doi.org/10.2139/ssrn.536942.

Kang, J., Kwon, K. Y. and Sim M. (2013), Retail Investor Sentiment and Stock Returns, The Korean Journal of Financial Management, 30(3), 35-68.

Kim, Y., Kim, N. and Jeong, S. R. (2012), Stock-Index Invest Model Using News Big Data Opinion Mining, Journal of Intelligent Information System, 18(2), 143-156.

Lee, D. H., Kim, S. H. and Kang, H. K. (2014), System Trading Using Big Data for KOSPI200 Futures, Proceeding of 2014 Conference on Korea Securities Association.

Park, W. J. (2012), Expectation and Concerns about Big Data Applications, Journal of Communications & Spectrum, 51, 28-47.

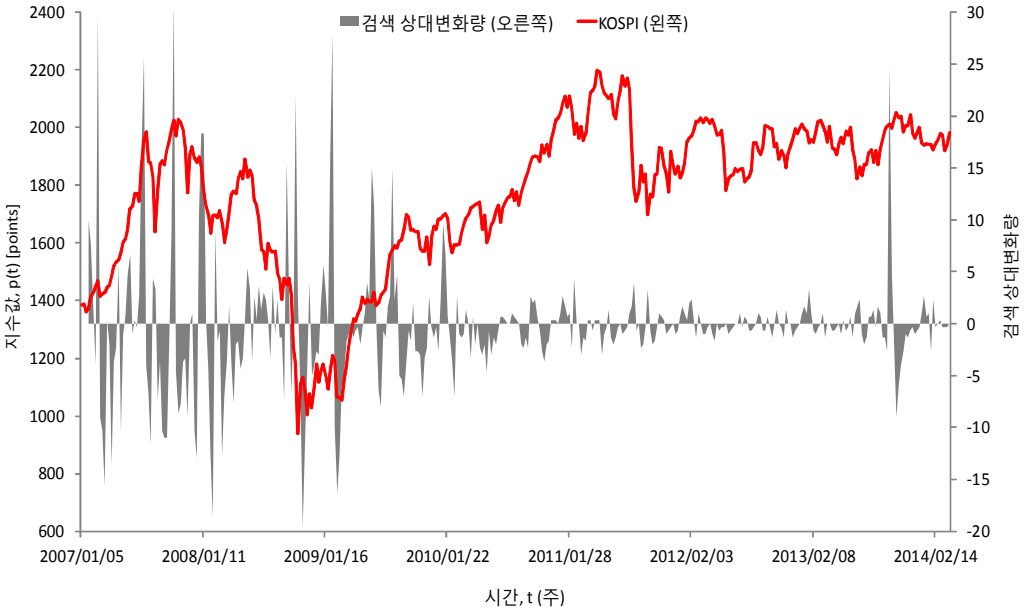
Preis, T., Reith, D. and Stanley, H. E. (2010), Complex Dynamics of Our Economic Life on Different Scales: Insights from Search Engine Query Data, Philosophical Transaction of the Royal Society A, 368, 5707-5719.

Preis, T., Moat, H. S. and Stanley, H. E. (2013), Quantifying Trading Behavior in Financial Markets Using Google Trends, Scientific Reports, 3, doi:10.1038/srep01684.

Toda, H. Y. and T. Yamamoto, “Statistical Inference in Vector Autoregressions with Possibly Integrated Processes”, Journal of Econometrics, 66, 1995, 225-250.

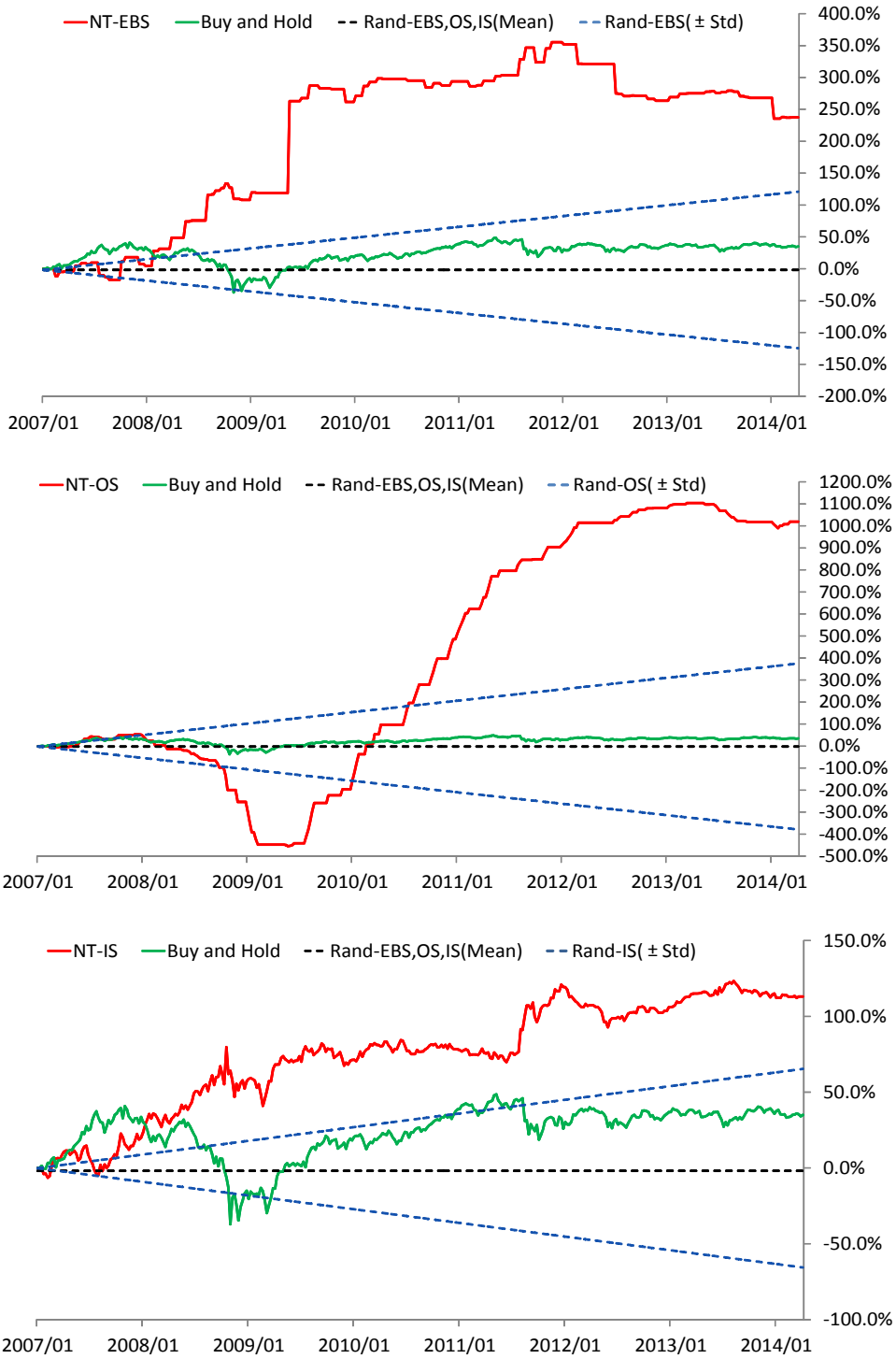
<그림 1> 검색량 변화에 따른 주식시장 움직임

2007년 1월 5일부터 2014년 3월 28일까지의 기간 동안 KOSPI 지수와 검색용어 CMA의 검색 상대변화량을 나타냄. 각각은 KOSPI 지수 t주 첫 거래일의 종가, 즉 $p(t)$의 시계열과 검색용어 CMA에 대해서 $\triangle t$ = 3주로 산출한 검색 상대변화량을 뜻한다.



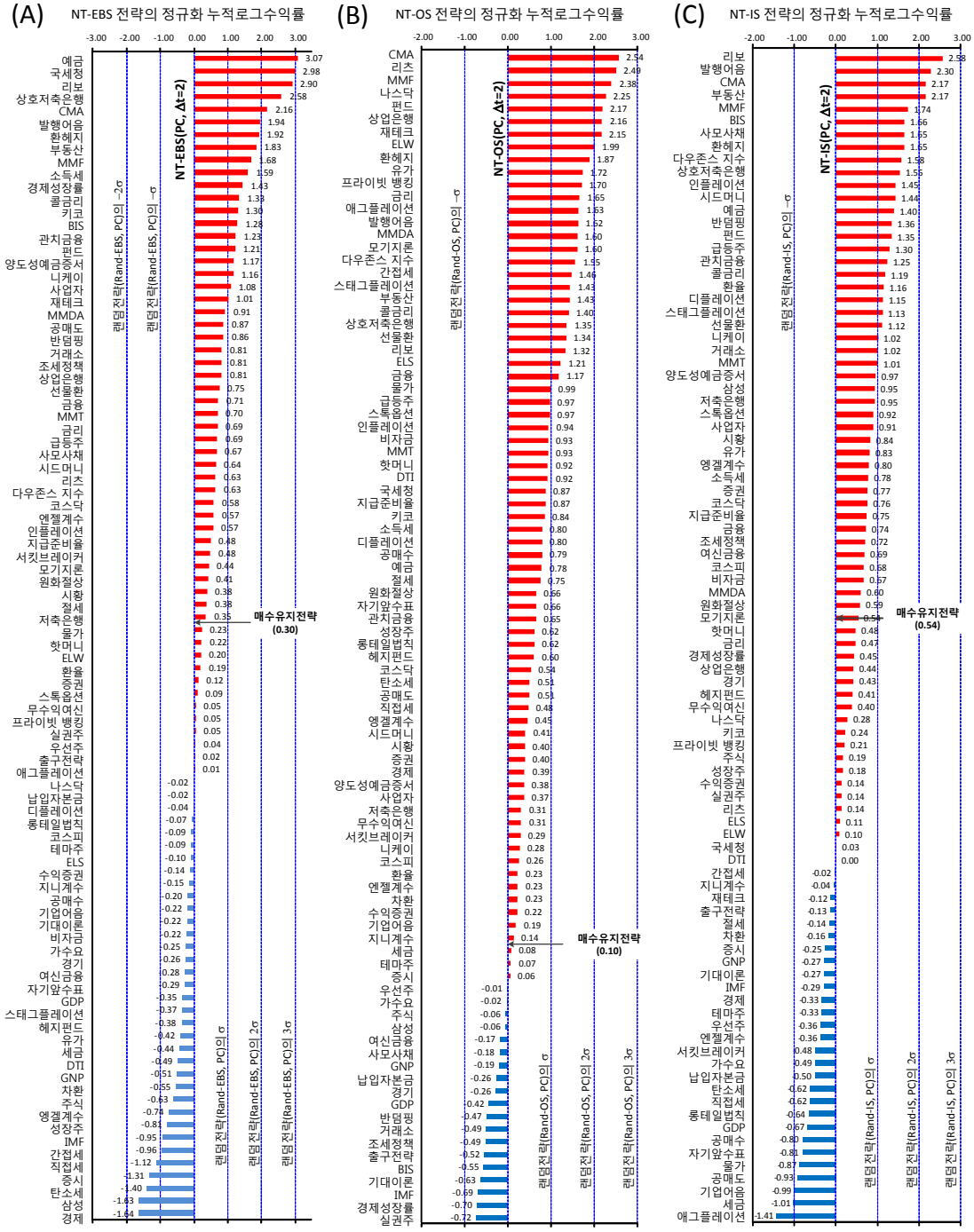
<그림 2> 네이버 트렌드 PC 검색어를 이용한 투자전략 간 누적수익률 비교

아래 차트는 차례대로 NT-EBS, NT-OS, NT-IS전략과 각각의 랜덤투자전략 및 매수유지(buy and hold) 전략들 간의 누적수익률 손익결과를 보여준다. 본 예시의 검색용어는 *CMA*이고, Δt =3이다. 매수유지전략은 세 차트에서 35.0%의 동일한 누적수익률을 나타낸다. 청산방법에 따라 랜덤투자전략도 Rand- EBS, Rand-OS, Rand-IS로 구분하며 괄호안의 Mean, +Std, -Std는 각각 10,000회 랜덤실험에서의 평균, 평균+표준편차, 평균-표준편차를 의미하고, 세 랜덤전략 모두 평균은 0에 근접함을 알 수 있다.



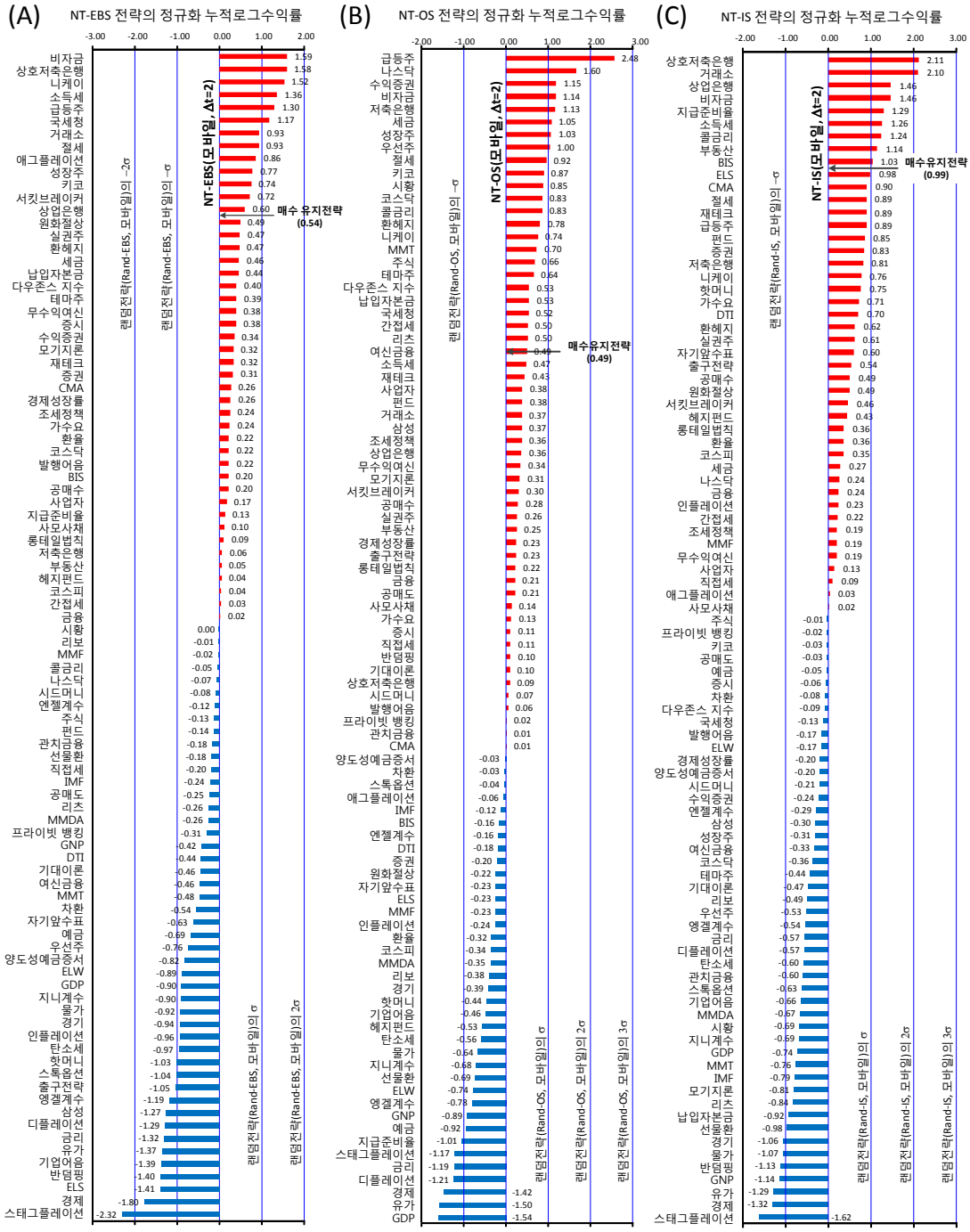
<그림 3> PC 기반 검색량 트렌드를 이용한 전략별 투자결과: (A) NT-EBS 전략, (B) NT-OS 전략, (C) NT-IS 전략

아래 차트는 2007년 1월부터 2014년 3월까지의 기간 동안 PC 기반의 트렌드를 반영하는 92개 검색어들의 검색량 데이터를 (A) NT-EBS, (B) NT-OS, (C) NT-IS 전략을 이용하여 투자한 결과를 보여준다. 양의 수익률은 빨간색 막대로, 음의 수익률은 파란색 막대로 구분하였고, 매수유지전략의 투자성과는 해당 차트에 각각 표시하였다. 각 차트의 막대 옆에 표기된 숫자는 정규화 누적로그수익률을 나타내며, NT-EBS 전략을 예로 들면 {(특정검색어의 NT-EBS전략 수익률 – Rand-EBS 수익률 평균) / Rand-EBS 수익률 표준편차}로 산출된다. 10,000회 시뮬레이션을 통해 얻은 Rand-EBS / Rand-OS / Rand-IS 랜덤전략의 수익률 평균은 0에 수렴하므로, 각 차트의 파란색 수직점선은 각각 랜덤전략 수익률 표준편차의 배수, 즉 -2σ, -σ, σ, 2σ, 3σ에 해당된다.



<그림 4> 모바일 기반 검색량 트렌드를 이용한 전략별 투자결과: (A) NT-EBS 전략, (B) NT-OS 전략, (C) NT-IS 전략

아래 차트는 2010년 7월부터 2014년 3월까지의 기간 동안 모바일 기반의 트렌드를 반영하는 92개 검색어들의 검색량 데이터를 (A) NT-EBS, (B) NT-OS, (C) NT-IS 전략을 이용하여 투자한 결과를 보여준다.



<표 1> 이동평균기간(Δt)에 따른 전략별 정규화 누적로그수익률 성과 비교

아래의 표는 제안한 전략들의 이동평균기간에 따른 정규화 누적수익률을 PC와 모바일로 구분하여 나타내고 있다. 각 이동평균기간의 숫자들은 전체 92개 검색어들의 정규화 누적로그수익률 평균을 의미하고, ALL 열은 전체 이동평균기간(Δt=1~6)의 평균이다. 랜덤투자전략과의 비교를 위해 단일표본 t-검정(귀무가설: 제안전략 수익률 평균 = 랜덤전략 수익률 평균, 대립가설: 제안전략 수익률 평균 > 랜덤전략 수익률 평균)을 시행하였다. 괄호안의 숫자는 t값이고 *, **, ***는 각각 p값 < 0.05, 0.01, 0.001일 경우에 해당한다. PC의 경우 전반적으로 세 전략모두 랜덤투자전략보다 매우 유의하게(p값 < 0.001) 높은 수익률을 보이고, 모바일의 경우 NT-OS전략만 유의하게 높은 수익률을 나타냄을 관찰할 수 있다.

매매전략	구분	이동평균기간(Δt, 주)						
		1	2	3	4	5	6	ALL
NT-EBS	PC	0.21 (3.43***)	0.38 (3.97***)	0.26 (2.29*)	0.25 (1.79*)	0.32 (1.99*)	0.34 (1.95*)	0.29 (5.52***)
	모바일	-0.1 (-1.61)	-0.11 (-1.37)	-0.15 (-1.37)	-0.22 (-1.86)	-0.34 (-2.58)	-0.44 (-2.98)	-0.23 (-4.94)
NT-OS	PC	0.37 (4.78***)	0.69 (8.47***)	0.86 (10.6***)	1.07 (11.0***)	1.27 (13.2***)	1.31 (11.7***)	0.93 (23.4***)
	모바일	0.05 (0.77)	0.09 (1.17)	0.09 (1.02)	0.05 (0.58)	0.08 (0.78)	0.08 (0.75)	0.07 (2.01*)
NT-IS	PC	0.23 (2.98**)	0.48 (5.48***)	0.34 (4.29***)	0.31 (4.14***)	0.28 (3.56***)	0.20 (2.31*)	0.30 (9.29***)
	모바일	0.06 (0.77)	0.05 (0.59)	-0.05 (-0.54)	-0.07 (-0.82)	-0.09 (-1.17)	-0.19 (-2.38)	-0.05 (-1.47)

<표 2> 시계열의 선행성과 투자성과 간의 상관관계 분석

본 표의 Panel A는 52개 검색어들의 검색량 시계열과 KOSPI 지수 시계열 간의 Granger인과관계 검증(귀무가설: 특정 검색어의 검색량 시계열이 KOSPI 지수 시계열을 Granger초래하지 않는다)을 통해 얻은 p값과 PC기반의 NT-EBS, NT-OS, NT-IS 전략들의 수익률 간의 Kendall rank correlation 검증 결과이다. Kendall rank correlation에 의하면 상관도를 의미하는 tau값이 전략별로 모두 통계적으로 유의한 음의 상관관계를 나타낸다. 이것은 검색량 시계열이 KOSPI 지수 시계열을 Granger초래할수록, 즉 선행성이 있을수록 높은 수익률을 산출한다는 것을 의미한다. *, **, ***는 각각 p-value < 0.1, 0.05, 0.01일 경우에 해당한다. Panel B는 Granger인과관계 검증의 수정된 왈드검정 결과인 p값의 구간별 평균 수익률을 보여주며, 각 구간의 대푯값을 각각 0.005, 0.025, 0.075, 0.3으로 설정했을 때의 Pearson 상관계수 값을 전략별로 나타낸다. 이 상관계수가 −1에 가까운 값을 갖는다는 것은 검색량 시계열이 KOSPI 지수 시계열에 대한 선행성이 높을수록 (수정된 왈드검정의 p값이 작을수록) 높은 수익률을 산출한다는 것을 뜻한다.

Panel A: Granger인과관계 분석에서 산출된 수정된 왈드검정의 p값과 전략별 투자수익률 간의 Kendall rank correlation 결과

Kendall statistics	NT-EBS	NT-OS	NT-IS	Average
tau	-0.128	-0.122	-0.19	-0.16
p-value	0.072*	0.085*	0.007***	0.024**

Panel B: 수정된 왈드검정 결과인 p값의 구간별 평균 수익률

Range of p-values	NT-EBS	NT-OS	NT-IS	Average
p-value < 0.01	130.91%	467.01%	84.12%	227.36%
0.01 ≤ p-value < 0.05	104.40%	400.53%	63.60%	189.52%
0.05 ≤ p-value < 0.1	68.67%	334.86%	47.86%	150.47%
p-value ≥ 0.1	33.58%	318.39%	23.69%	125.22%
Pearson's r	-0.91	-0.76	-0.91	-0.84

Investment strategies for KOSPI Index using Big Data Trends

Hyun Joon Shin
Ra HyunWoo

<Abstract>

This study recognizes that there is a correlation between the movement of the financial market and the sentimental changes of the public participating directly or indirectly in the market, and applies the relationship to investment strategies for stock market. The concerns that market participants have about the economy can be transformed to the search terms that internet users query on search engines, and search volume of a specific term over time can be understood as the economic trend of big data. Under the hypothesis that the time when the economic concerns start increasing precedes the decline in the stock market price and vice versa, this study proposes three investment strategies using casuality between price of domestic stock market and search volume from Naver trends, and verifies the hypothesis. The computational results illustrate the potential that combining extensive behavioral data sets offers for a better understanding of collective human behavior in domestic stock market.

Keywords: Big Data, Trends, Market Sentiment, Search Terms, Query Volume, Investment Strategies