정보화연구
제13권 1호

ISSN : 1738-382X(Print)

# 빅데이터 트렌드를 이용한 섹터 투자 전략

유재필, 한창훈, 신현준

To cite this article : 유재필, 한창훈, 신현준 (2016) 빅데이터 트렌드를 이용한 섹터 투자 전략, 정보화연구, 13:1, 111-121

Journal of Information Technology and Architecture
Vol. 13. No. 1, March 2016, Pages 111-121

# 빅데이터 트렌드를 이용한 섹터 투자 전략

유재필[1] · 한창훈[2] · 신현준[2*]
[1]KIS채권평가 금융공학연구소
[2]상명대 경영공학과
jaepilryu@kispricing.com; hahn65@gmail.com; hjshin@smu.ac.kr

(2016년 3월 6일 접수; 2016년 3월 22일 수정; 2016년 3월 28일 채택)

**요약:** 빅데이터를 다양한 산업분야에 적용하려는 시도가 증가하고 있는 가운데 빅데이터 트렌드를 금융시장에 활용하려는 연구들이 활발히 진행되고 있다. 더불어 금융시장 관련 빅데이터 트렌드가 주식시장의 움직임을 선 반영할 수 있다는 사실이 최근 연구들에 의해 입증되고 있다. 기존 연구와 달리 본 연구에서는 주식 시장의 트렌드를 보다 세분화하여 포착하기 위해 주식 시장을 11개의 섹터로 세분화한다. 각 섹터의 트렌드를 대표하는 키워드들을 텍스트마이닝과 브레인스토밍 기법을 통해 각각 선정하고 5년간의 트렌드 데이터를 수집함으로써 섹터별 상장지수펀드(ETF) 투자 포트폴리오 전략을 수립한다. 섹터별 투자성과를 누적수익률 및 연도별 수익률 관점에서 비교한 결과 텍스트마이닝 기법에 기반을 둔 섹터 트렌드 전략이 보다 우수한 성과를 보이는 것으로 나타났다.

**키워드:** 빅데이터, 트렌드, 텍스트마이닝, 섹터투자, 상장지수펀드, 포트폴리오 투자 전략

# Sector Investment strategies Using Big Data Trends

Ryu Jaepil[1], Hahn Chang Hoon[2], and Shin Hyun Joon[2*]
[1]Center for Financial Engineering, KIS Pricing
[2]Department of Management Engineering, Sangmyung University

(Received March 6, 2016; Revised March 22, 2016; Accepted March 28, 2016)

**Abstract:** Recently, researches on applying big data trends to financial market have been actively conducted, while a lot of attempts using big data for various industries are increasing. In addition, researches show 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. Unlike previous studies, this study breaks down the stock market into 11 sectors in order to closely capture the trends from each markets. Keywords for each sector are selected by text mining and brainstorming methods, and trends data of these keywords are collected for recent five years. The computational results illustrate that the invest strategy based on text mining shows better performance than one based on brain storming in terms of accumulated rate of returns.

**Keywords:** Big Data, Trends, Text Mining, Sector Investment, ETF, Portfolio Investment Strategies

## 1. 서 론

구글 트렌드(Google Trend)는 검색어의 빈도를 통해서 해당 단어(keyword)에 대한 세계적인 관심도를 시계열로 보여주고 있다. 빅데이터 단어에 대한 검색 빈도는 2011년부터 큰 폭으로 증가하고 있는데 100을

기준으로 인도가 100으로 가장 높고 한국이 78로 세계 2위이며 그 외 미국은 47, 일본은 8로 매우 낮다. 특정 주제에 대해서 사회 여론화가 매우 빠르게 나타나는 한국의 문화 적 특성을 감안하더라도 빅데이터에 대한 국내 관심도는 매우 높은 수준이다. 또한 스마트폰의 보급 확산으로 인해서 포털과 같은 인터넷상에서의 데이터 축적은 급속도로 증가하고 있으며, 2020년에는 약 40제타바이트의 시대가 올 것이라고 예상한다(김정래, 정찬기, 2013; 임유진, 최은미, 2015). 이처럼 빅데이터는 민간 및 공공부문에 있어서 다양한 부가가치를 창출하고 효과적인 활용으로 생산성 향상, 기업의 경쟁력 제고 그리고 국가 미래전략 지원 및 공공 서비스의 혁신이라는 기대감을 주고 있다. 이처럼 빅데이터가 주목받는 가장 주된 요인은 대용량 데이터의 분석과 추론을 통해서 새로운 서비스를 제공할 수 있고 금융업, 제조업, 유통업 등 다양한 산업에 적용가능하다는 것이다. 또한 온라인 산업과 서비스가 발달하면서 온라인 네트워크 이용자들의 트렌드(trend)에 대한 분석이 빅데이터 분석을 통해서 가능하고 이는 기업에서 다양한 의사결정을 하는데 있어서 중요한 척도가 되고 있다. 특히 포털 빅데이터는 온라인 네트워크를 이용하는 사람들의 일상생활에 내재된 행동양식이 반영되어 있으며 복잡한 현상에 대한 근본적인 문제를 해결하기 위한 새로운 기회가 된다. 이러한 포털 빅데이터는 주식 시장과도 밀접한 관련이 있다.

주식을 투자하는 사람들은 대부분 포털을 통해서 다양한 투자 정보를 수집한다. 그리고 주가와 관련이 깊은 다양한 경제적 용어 및 투자 주체에 대한 연관된 단어들을 포털을 통해서 검색을 한다. 이미 야후(Yahoo)에서 제공하는 트렌드를 통해서 주식 시장의 거래량과 이와 관련된 단어들의 검색량 간의 연관성이 있다는 것과 구글 트렌드 데이터가 자동차 판매량, 실업수당 신청률, 소비자신뢰지수 등 여러 경제적 지표와 연계될 수 있다는 것이 연구를 통해서 입증되었다(Choi and Varian, 2012). 또한 투자자는 시장에 대한 불안감이 증폭되는 기간 동안에는 주식 매입 또는 매도에 대한 의사결정을 위해서 시장에 대한 다양한 정보 검색을 평소보다 더 많이 한다는 것을 구글 트렌드 분석을 통해서 밝혀진바 있다(Preis T, Moat H S and Stanley H E, 2013). 이러한 현상은 투자자의 심리와 밀접한 관련이 있다. 실제로 일반 개인 투자자들은 정량적인 트레이딩 시스템(trading system)과는 거리가 멀고 대부분 포털 검색을 통해서 정보를 수집한 후 정성적인 판단을 통해서 매매하는 경향이 있기 때문이다.

이처럼 투자자의 투자 심리는 주식 시장과의 밀접한 연관이 있는데 이는 투자자의 심리적 요인을 분석하여 투자 전략을 수립하는 방식과 함께 빅데이터 등의 트렌드를 정량적으로 분석하여 투자 전략을 수립하는 배경이 되고 있다. 연구를 통해서도 미국의 주식 시장에서 개인 투자자는 집단적 거래행태가 존재하고 개인 투자자의 투자 심리와 주식 수익률 간의 유의한 관계가 있음을 보였다(Kumar and Lee, 2006). 그 외 호주와 독일의 주식 시장에서도 브로커를 통한 개인 투자자들의 거래를 각각 분석함으로써 개인 투자자들 간의 거래에 있어서 유의한 상관관계가 있다는 것이 입증되었다(Jackson A, 2003; Dorn D, Huberman G and Sengmueller P, 2008). 이와 같이 투자 심리에 대한 연구는 빅데이터나 매체를 통해서 투자자의 심리를 분석하는 연구에 비해서 많이 연구되어 왔다. 그리고 주식 투자자들이 의견을 공유할 수 있는 포털과 SNS(Social Networking Service) 등이 대중화되면서 주식 투자를 하는 투자자들의 약 93%가 포털을 통해서 투자 주체 및 그와 연관된 단어 검색을 하고 있으며, 최근 빅데이터의 관심도가 증가하면서 이와 관련된 연구도 함께 진행되고 있다. 따라서 본 연구에서는 주식시장을 세분화하여 분할된 섹터 내에서 수집된 트렌드를 해당 섹터에 반영하여 매매하는 일종의 섹터투자 전략을 제시하고자 한다. 이를 위해서 주식시장을 11개의 섹터로 구분하고 각 섹터별 35~50개의 트렌드 검색어들을 텍스트마이닝과 브레인스토밍 기법을 이용하여 추출하여 섹터투자 전략 포트폴리오를 구성한다. 매매는 11개 섹터별로 거래되고 있는 ETF(exchange traded fund) 상품들을 대상으로 한다. 4년간(2011 ~ 2014)의 섹터별 투자 성과를 누적수익률 관점에서 비교 분석함으로써 유의미한 결과를 도출하고자 한다.

## 2. 관련문헌 연구

Barber B M, Odean T and Zhu N(2009)은 감정이 개개인의 행동과 의사결정에 큰 영향을 미친다는 행동경제학(behavioral economics) 이론을 바탕으로 대규모 트위터 피드(twitter feeds)로부터 추출한 집단적 감정 상태의 변화가 일정한 기간 동안에 다우존스산업지수(DJIA)의 가치 변화와 상관관계가 있음을 밝혔다. 또한 2004년부터 2010년까지 구글 검색 엔진을 통해서 질의된 다양한 검색어들의 주별(weekly) 검색량과 주식 시장

의 변동성간에 상관관계가 있다는 점과 특히 S&P500에 포함된 개별 종목들의 검색량이 개별주식의 거래량과도 유의한 상관관계가 있다는 것이 연구결과 입증되었다 (Preis T, Reith D and Stanley H E, 2010). Bordino et al(2012)은 2010년 5월부터 2011년 4월까지 야후의 검색 엔진을 통해서 질의된 검색어의 검색량과 NADAQ100에 상장된 종목들의 거래량 간의 상관관계를 보였고, 이들 두 시계열 간의 인과관계를 분석하였다. 김유신(2012)은 오피니언 마이닝을 통해서 지능형 투자의사결정모형을 제시하였고 이를 통해서 주가지수 변동성을 예측하는 가능성을 제시하였다. 박원준(2012)은 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. 이득환(2014)은 빅데이터에 나타난 투자자별 감성이나 정보가 KOSPI200 선물 지수 수익률에 미치는 영향을 실증분석하여 빅데이터가 KOSPI200 선물 지수 수익률을 예측하는 정보를 포함한다는 것과 빅데이터를 사용한 투자 전략이 높은 수익을 가져옴을 증명하였다.

앞서 설명한 기존의 연구를 살펴보면 우리나라 인터넷 포털 검색엔진에서 형성되는 섹터별 빅데이터 트렌드를 국내 주식시장 섹터투자전략에 적용한 연구는 찾아보기 어렵다. 따라서 본 연구는 국내 대표적인 검색엔진을 통해 형성되는 사용자의 경제관심의 변화 트렌드를 섹터별 ETF(Exchange Traded Funds)에 반영하는 투자전략을 제안하고 그 성과를 분석하고자 한다. 일반적으로 시장에 대한 불안감이 높아질수록 시장에 대한 관심도는 증가한다(Preis T, Moat H S and Stanley H E, 2013). 예컨대 투자자가 투자한 자산에 대한 불안감이 높아지면 포털에 투자 자산과 관련된 단어를 검색하거나 SNS를 통해서 본인의 불안한 심리를 표현한다. 또한 증권관련 포털을 보면 종목 상담의 대부분은 현재의 주가가 매입가보다 높을 경우에 비해서 낮을 경우가 월등하게 많다. 신현준(2015)은 국내 빅데이터 트렌드를 이용하여 KOSPI 주가지수 투자 전략을 수립하고 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 것을 입증하였다. 따라서 본 연구에서는 국내 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 국내 주식시장 주가의 하락 시점을 선행하며 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 가정 하에, KOSPI 주가지수에 일관된 투자 주체를 넘어서 상장지수펀드인 ETF에 본 연구에서 제안하는 매매 전략을 제안한다. 이를 위해서 투자 대상의 경제 전반에 대한 시장참여자의 심리(sentiment)를 대표하는 50개의 트렌드 검색어를 텍스트마이닝(Text Mining) 기법을 통해서 선정한다. Figure 1은 텍스트마이닝을 통해서 단어를 추출하는 과정과 ETF 매매과정을 도식화한 그림인데 자세한 설명은 다음 장에 기술한다.

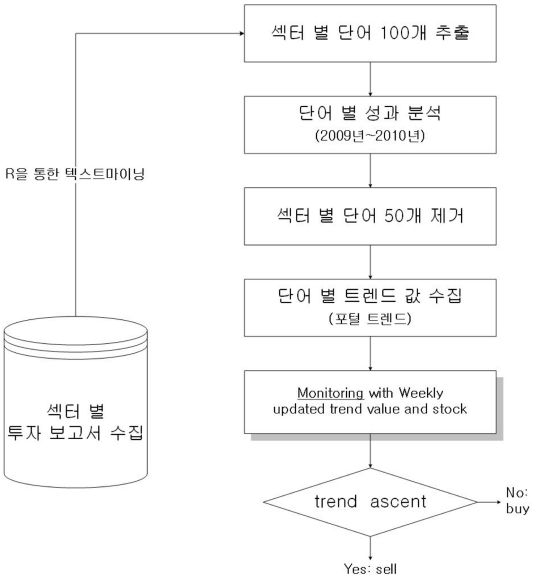


Figure 1. 텍스트마이닝을 통한 ETF 매매 과정

본 논문의 3장에서는 실험을 위한 매매 대상인 섹터 선정과 선정된 섹터와 연관된 단어를 선정하는 과정에 대해서 설명한다. 4장에서는 섹터 별로 연관된 단어를 바탕으로 섹터 투자 전략에 대해서 설명하고, 5장에서는 본 연구의 실험 계획 및 실험 결과 그리고 본 연구에서 제안하는 섹터 투자 전략에 대한 운영 성과 평가 등에 대해서 설명한다. 마지막으로 6장에서는 결론으로 마무리 한다.

## 3. 연구의 자료

본 연구에서는 포털 트렌드의 변화량을 이용하여 다양한 ETF에 투자하는 전략을 제안한다. 따라서 투자 대상과 투자 대상과 관련된 단어를 선정하는 과정은 본 장에서 설명하고자 한다.

### 3.1 섹터 선정

일반적으로 개별 종목에 대한 포트폴리오를 구성하는 것은 너무 많은 수의 종목들이 투자 대상에 포함되기 때문에 종목을 선정하는 것이 힘들며, 몇 몇의 투자자로 인해서 명확한 근거가 없이 주가가 급등 또는 급락하는 종목들이 있기 때문에 효과적인 포트폴리오를 구성하는 전략을 수립하는 것이 매우 어렵다. 반면 상장지수펀드인 ETF(Exchange Traded Funds)는 KOSPI200 또는 KOSPI50과 같이 다소 우량한 기업들을 구성하고 인덱스 펀드와는 달리 거래소에 상장되어 일반 주식처럼 자유롭게 사고 팔 수 있기 때문에 본 연구와 같이 새로운 투자 전략을 제안하고자 실험하는 연구에 있어서 적합하다고 사료된다. 따라서 본 연구는 한국 거래소에 상장된 총 11개의 섹터인 TIGER 은행, KODEX 에너지화학, KODEX 운송, KODEX 조선, KODEX 반도체, KODEX 철강, TIGER IT, KODEX 소비재, KODEX 건설, KODEX 자동차, TIGER 미디어 통신을 선정한다. 또한 포털에서 제공하는 트렌드의 변화량 자료가 주 별로 제공된다는 점을 감안하여 각 섹터의 주가를 증권사에서 제공하는 HTS(Home Trading System)를 통해서 본 연구의 실험 기간인 2009년부터 2014년까지 주 별로 수집한다. ETF는 일반적으로 KOSPI 지수와 양의 상관관계를 보이는데 본 연구의 투자 대상인 11개의 섹터와 KOSPI 지수와의 상관계수는 Table 1과 같다. KOSPI 지수와 섹터간의 상관계수를 보면 TIGER 은행이 KOSPI 지수와 가장 높은 상관관계를 보인다. 이는 은행들의 주가는 큰 변동성이 없이 KOSPI 지수와 유사한 방향으로 움직이기 때문이다. 다음으로 반도체가 0.60으로 높은데 KODEX 반도체에 KOSPI 지수의 움직임에 높은 영향을 주는 삼성전자가 구성되어 있기 때문이다. 다음 절에서는 각 섹터 별로 키워드(Keyword)를 선정하는 과정과 키워드 선정을 위해 사용하는 텍스트마이닝에 대

Table 1. 섹터와 KOSPI 지수와의 상관계수

| 섹터종류 | IT | 건설 | 미디어 | 반도체 | 소비재 | 에너지 | 운송 | 은행 | 자동차 | 조선 | 철강 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 상관계수 | 0.50 | 0.20 | 0.22 | 0.60 | 0.32 | 0.40 | 0.42 | 0.77 | 0.37 | 0.44 | 0.43 |

Table 2. 최종 선정된 에너지화학 섹터 키워드

| 최종적인 에너지화학 섹터 키워드 | | | | |
|---|---|---|---|---|
| S-OIL | 두바이유 | 신재생에너지 | 연료 전지 | 한국선급 |
| LG화학 | SK가스 | 프로필렌 | 합성수지 | 부타디엔 |
| Opex | bbl | LNG | 배럴 | 브렌트유 |
| 한샘 | 현대리바트 | 라이온켐텍 | KCC | 호남석유화학 |
| 국제유가 | SK이노베이션 | 한화케미칼 | 동남아 | 친환경 |
| 천연가스 | 원자력 | 휘발유 | 석유 | 폴리우레탄 |
| 톨루엔 | HDPE | MEG | PVC | BPA |
| 스프레드 | 부동산 가격 | 플라스틱 | 나프타 | 에틸렌 |
| 원유 | 정유 | 석유화학 | 유가 | 석탄 |
| ABS | LPG | BENZENE | 디젤 | 태양전지 |
| 제거된 에너지화학 섹터 키워드 | | | | |
| 경기 | 코스닥 | 테마주 | 삼성 | 예금 |
| 조세정책 | 코스피 | 시황 | 헤지펀드 | 스톡옵션 |
| 간접세 | 기업어음 | 급등주 | 출구전략 | 중국 |
| 직접세 | 콜금리 | ELW | 경제 | 국민은행 |
| 환헤지 | 우선주 | 금융 | 모기지론 | 신용등급 |
| 나스닥 | 주식 | 유가 | 공매도 | 정책 |
| 지급준비율 | 재테크 | 거래소 | 경제성장률 | 부양 |
| GDP | 증권 | 증시 | 인플레이션 | 경매 |
| GNP | 부동산 | ELS | 물가 | 서울은행 |
| 환율 | 펀드 | 탄소세 | 세금 | 중장비 |

해서 설명하고자 한다.

### 3.2 텍스트마이닝

텍스트마이닝은 비정형화된 대규모 텍스트 자료로부터 일반화된 패턴 및 관계를 발견하고 추출하는 과정이다(강병욱, 2015). 텍스트마이닝은 크게 데이터를 수집하는 과정, 용어를 추출하는 과정, 정보를 추출하는 과정 그리고 정보를 분석하는 과정의 총 4단계의 과정을 거친다. 먼저 텍스트마이닝의 첫 번째 단계인 데이터를 수집하는 과정은 비정형 대규모 텍스트 자료를 수집하는 단계이다. 두 번째 용어를 추출하는 과정은 구문의 패턴이나 단어들의 연관성을 고려하여 연관성 분석에 의해 단어들을 추출하여 분석 대상 용어들의 후보를 선정하고, 선정된 단어에 대해서 여러 가지 통계적인 방법을 통해서 전체를 대표하는 단어들을 추출하는 과정이다. 즉 첫 번째 단계에서 얻어진 분석대상 데이터를 중심으로 정보를 추출할 수 있도록 데이터를 가공하는 단계로 수집된 문서를 기본으로 관련 키워드를 추출한다. 세 번째 정보를 추출하는 과정은 문서 자체를 찾는 것이 아니라 문서 내에서 유용한 정보를 찾는 과정이다. 예컨대 제품을 소개하는 문서의 경우에 제품명, 제품의 기능 그리고 주의 사항 등과 같이 상세 정보를 포함하는 용어를 추출하기 위한 과정이다. 네 번째 정보를 분석하는 과정은 세 번째 과정에서 얻어진 최종 단어에 대해서 빈도, 분류, 군집화(Clustering) 그리고 컨셉 도출 방법 등을 이용하여 유용한 정보를 도출해 내는 과정이다. 빈도는 가장 많이 얻어지는 단어가 무엇인지에 대한 정보를 도출해 내고, 분류는 앞서 추출된 단어의 내용에 따라 문서들을 범주화 시켜주는 과정이다. 즉 주어진 신문 텍스트 문서가 금융 분야인지 또는 정치 분야인지 등을 단어에 따라 분류하는 것을 의미한다. 군집화는 문서에 포함되어 있는 추출된 단어들을 유사도에 따라 여러 개의 텍스트(Text) 집단으로 군집화 시켜주는 과정이다. 컨셉도출은 어떤 특정한 키워드를 중심으로 또 다른 키워드들 간의 관계를 파악하는 기법이다.

본 연구에서는 섹터별로 25개의 리포트를 선정하는데 선정하는 기준은 2014년도에 매출이 가장 높았던 3개의 증권사에서 발행하는 리포트를 대상으로 한다. 또한 3개의 증권사에서 발행하는 리포트가 방대하기 때문에 해당 섹터와 동일한 주제의 리포트 중에서 조회


Figure 2. R을 통해 선정된 에너지화학 섹터 키워드

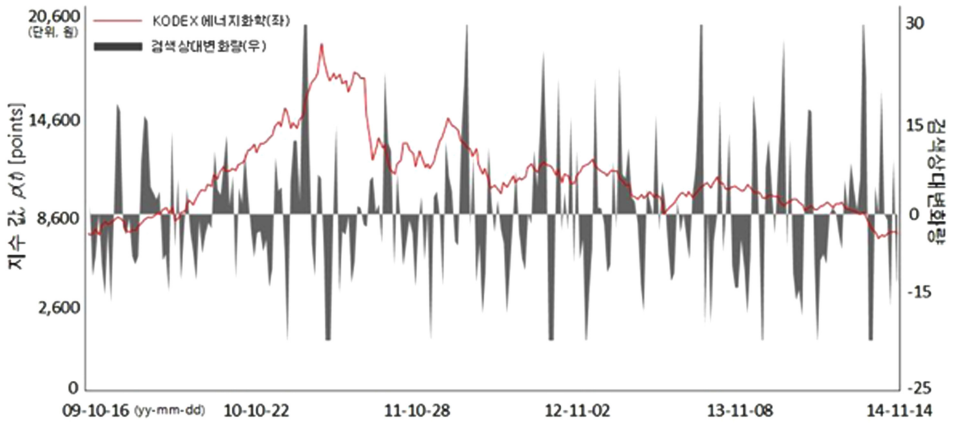


Figure 3. 에너지화학 종가와 석유화학 키워드간의 검색 상대변화량

수가 가장 많은 순으로 리포트를 선정한다. 선정된 리포트를 통해서 섹터 별로 해당 키워드를 선정해야 하는데 이는 다양한 분야에서 사용되는 프로그램 R을 이용한다. 텍스트마이닝을 구현하는 과정은 tm 패키지를 이용하여 문서별 Corpus 생성, 명사 추출, 불 용어 및 기타 의미 없는 기호를 제거, 두 자리 이상의 명사 추출 그리고 KoNLP 패키지를 사용하여 Corpus 내의 한글 형태소 단위를 인식하는 절차로 구성된다. 그리고 명사들의 빈도를 이용하여 Matrix 구조의 유사도를 판단 및 구현하고 유사한 문서의 그룹화를 거쳐서 키워드와 복합 명사들을 추출한다(이종화, 2015). Figure 2는 KODEX 에너지화학에 해당하는 키워드를 R을 통해서 추출한 결과를 사례로 보여주는 그림이며, 해당 섹터에 대한 대표성이 강한 단어가 더 크게 보인다. 25개의 리포트에서 R을 통해서 100개의 단어를 추출하고 2009년부터 2010년까지 각 단어의 트렌드 변화량을 바탕으로 섹터 투자를 수행한 후, 이 중에서 성과가 높았던 50개를 최종적으로 해당 섹터의 연관 단어로 선정한다. Table 2는 KODEX 에너지화학에 해당하는 최종 키워드들이다.

## 4. 섹터 투자 전략

앞서 설명한 섹터 선정과 해당 섹터의 키워드를 텍스트마이닝을 통해서 선정하면, 다음으로는 투자 전략을 수립해야 한다. 텍스트마이닝을 통해서 선정된 키워드의 트렌드는 네이버 트렌드를 통해서 자료를 수집하는데 이는 주 별 자료를 수집한다. 따라서 전 주 대비하여 해당 키워드의 검색량이 증가하면 해당 섹터의 ETF는 매도하는 전략을 취한다. 예컨대 KODEX 에너지화학의 키워드로 선정된 석유화학 키워드가 전 주 대비하여 검색량이 떨어지면 KODEX 에너지화학을 매수한다. 분석의 단순화를 위해 투자 대상인 ETF는 공매도(short selling)가 가능하다고 가정한다. 또한 키워드의 변화량을 전 주 대비가 아닌 과거 2주, 3주 등의 이동평균 값의 변화량을 바탕으로 매매 실험을 진행한다. 예컨대 $t-1$ 주에 MMF 용어가 네이버에서 검색된 검색량, $n(t-1)$ 을 네이버 트렌드를 통해 산출할 수 있다. 시장 참여자들의 정보검색 움직임을 정량화하기 위해 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 $\{n(t-1)+n(t-2)+\ldots+n(t-\Delta t)\}/\Delta t$ 로 정의되며 $t-1$ 주부터 최근 $\Delta t$ 주 동안의 검색량의 이동평균을 뜻한다 (Preis T, Moat H S and Stanley H E, 2013). Figure 3은 KODEX 에너지화학의 주별 종가와 검색용어 석유화학에 대한 검색 상대변화량의 시간에 따른 변화를 보여주며, 각각은 KODEX 에너지화학 $t$주 첫 거래일의 종가, 즉 $p(t)$ 의 시계열과 검색용어 석유화학의 $\Delta t$ = 3주로 산출한 검색 상대변화량을 뜻한다.

실험하는 과정에서 매매 전략은 신현준(2015)의 연구에서 제안한 NT-OS전략을 바탕으로 실험한다. 일대일청산(one-to-one settlement; 이하 OS)을 의미하는 NT-OS전략은 매매신호 발생 시에 취한 포지션을 바로 청산하지 않고 보유해 나가는 방식으로써 청산은 기존의 유지하고 있는 포지션과 반대의 매매신호가 발생하면 보유 중인 가장 오래된 반대 포지션과 일대일로 시행한다. 예컨대 만약 $t-1$ 주에 $\Delta n(t-1,\Delta t) > 0$ 이라면 $t$주의 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 $t$주의 매도포지션과 일대일로 청산된다. 이 경우 누적수익률은 $\log p(t) - \log p(t+3)$ 이다. NT-OS전략은 수익률을 극대화할 수 있는 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.

## 5. 실험결과 및 분석

### 5.1 실험계획

본 연구에서 제안하는 섹터 투자 전략의 성능을 분석하기 위해서 Table 3의 실험계획을 수립한다. 앞서 설명했듯이 투자 대상은 총 11개의 ETF이며, 매매 전략은 앞 장에서 설명한 NT-IS전략을 이용한다. 매매 시 발행하는 거래 수수료는 0.004%로 정의하며 본 연구의 실험 기간은 2011년부터 2014년까지 설정한다. 또한 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매한 결과와 비교하기 위해서 추가적으로 50개의 키워드를 선정하고 이를 이용하여 매매한 결과와 비교한다. 이는 금융업에 종사하는 5명을 통해서 각 섹터 별로 의미 있는 키워드를 선정하였다. 즉 정량적인 텍스트마이닝 기법을 통해서 추출된 키워드들을 바탕으로

Table 3. 실험 계획

| 실험 요인 | 값 |
|---|---|
| 투자 대상 | 11개의 ETF(Exchange Traded Funds) |
| 키워드 선정 방법론 | 텍스트마이닝(Text Mining) |
| 거래 수수료 | 0.004% |
| 비교대상 | BS-T 기법 |
| 실험 기간 | 4 년(2011.01.02~2014.12.26) |

ETF를 매매한 결과와 정성적인 판단에 의해서 추출한 키워드들을 바탕으로 ETF를 매매한 결과를 비교 분석함으로써 본 연구에서 제안하는 전자의 방법론이 어느 정도의 성과가 있는지를 보기 위함이다. 본 방법론은 편의상 브레인스토밍(Brain Storming; 이하 BS)기법이라고 정의한다. 더불어 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매(Trading)하는 방식을 TM-T으로 BS 기법을 통해서 선정된 키워드를 바탕으로 매매하는 방식을 BS-T라고 정의한다.

### 5.2 실험결과 및 분석

Table 4는 TM-T와 BS-T를 바탕으로 섹터 별 매매한 연도별 로그수익률과 누적 로그수익률에 대한 결과를 정리한 표이다. 표에서 $\Delta 1$, $\Delta 2$, $\Delta 3$ 은 이동평균기간($\Delta t$, 주)를 의미한다. 대체적으로 텍스트마이닝을 통해서 추출한 단어를 바탕으로 매매한 TM-T가 수익률

Table 4. 매매 방법론에 따른 섹터 별 수익률

| | | 2011 | | | 2012 | | | 2013 | | | 2014 | | | 누적수익률 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Δ1 | Δ2 | Δ3 | Δ1 | Δ2 | Δ3 | Δ1 | Δ2 | Δ3 | Δ1 | Δ2 | Δ3 | Δ1 | Δ2 | Δ3 |
| 건설 | TM-T | -0.10 | -0.12 | -0.13 | 0.40 | 0.48 | 0.46 | 0.68 | 0.77 | 0.82 | 1.26 | 1.28 | 1.41 | 3.81 | 4.24 | 4.59 |
| | BS-T | 0.01 | 0.44 | 0.41 | -0.07 | 0.30 | 0.54 | 0.09 | 0.22 | 0.27 | 0.22 | 0.29 | 0.44 | 0.23 | 1.97 | 2.98 |
| 소비재 | TM-T | -0.07 | -0.27 | -0.43 | 0.09 | 0.14 | 0.07 | 0.22 | 0.30 | 0.36 | 0.88 | 0.96 | 0.84 | 1.31 | 1.09 | 0.52 |
| | BS-T | -0.10 | -0.30 | -0.40 | 0.07 | 0.13 | 0.11 | 0.19 | 0.18 | 0.21 | 0.75 | 0.92 | 1.05 | 0.99 | 0.80 | 0.67 |
| 에너지 | TM-T | 0.32 | 0.34 | 0.23 | 0.31 | 0.38 | 0.41 | 0.58 | 0.61 | 0.76 | 0.21 | 0.33 | 0.33 | 2.30 | 2.97 | 3.04 |
| | BS-T | 0.17 | 0.10 | 0.07 | 0.51 | 0.75 | 0.51 | 0.47 | 0.58 | 0.18 | 0.77 | 0.67 | 0.81 | 3.62 | 4.08 | 2.46 |
| 자동차 | TM-T | 0.55 | 0.71 | 0.48 | 0.19 | 0.28 | 0.60 | -0.17 | 0.08 | 0.09 | 0.20 | -0.07 | -0.07 | 0.84 | 1.21 | 1.42 |
| | BS-T | -0.04 | -0.28 | -0.61 | 0.18 | 0.20 | 0.27 | -0.12 | 0.15 | 0.15 | 0.14 | -0.17 | -0.14 | 0.12 | -0.18 | -0.51 |
| 조선 | TM-T | 0.24 | 0.38 | 0.18 | 0.29 | 0.27 | 0.28 | -0.04 | 0.55 | 0.07 | 0.24 | 0.11 | 0.68 | 0.90 | 2.01 | 1.72 |
| | BS-T | -0.92 | -0.39 | 0.39 | -0.88 | -1.81 | -2.84 | 0.13 | -0.31 | -0.62 | -2.85 | -2.98 | -3.27 | -1.02 | -0.32 | 1.23 |
| 철강 | TM-T | 0.11 | 0.26 | 0.17 | -0.04 | 0.12 | -0.08 | 0.33 | 0.02 | 0.13 | 0.26 | 0.12 | 0.03 | 0.78 | 0.61 | 0.25 |
| | BS-T | -0.17 | -0.19 | -0.59 | -0.04 | -0.18 | 0.24 | 0.08 | -0.38 | -0.31 | 0.00 | 0.19 | 0.11 | -0.13 | -0.51 | -0.61 |
| 반도체 | TM-T | -0.33 | 0.12 | 0.09 | 0.15 | 0.02 | -0.19 | 0.34 | 0.46 | 0.58 | 0.40 | 0.56 | 0.66 | 0.45 | 1.60 | 1.29 |
| | BS-T | -0.97 | -1.22 | -1.08 | 0.06 | 0.27 | 0.19 | 1.20 | 1.22 | 1.24 | 0.77 | 0.78 | 0.85 | -0.88 | -2.10 | -1.38 |
| IT | TM-T | 0.06 | 0.09 | 0.24 | 0.13 | 0.15 | 0.22 | 0.44 | 0.47 | 0.60 | 0.73 | 0.65 | 0.74 | 1.98 | 2.03 | 3.22 |
| | BS-T | 0.31 | 0.32 | 0.41 | -0.38 | -0.54 | -0.60 | -0.28 | -0.35 | -0.49 | -0.28 | -0.40 | -0.57 | -0.58 | -0.76 | -0.88 |
| 미디어 | TM-T | 0.24 | 0.45 | 0.61 | 0.02 | 0.06 | 0.14 | 0.55 | 0.12 | 0.06 | 0.08 | 0.32 | 0.74 | 1.11 | 1.27 | 2.36 |
| | BS-T | -0.33 | -0.41 | -0.45 | -0.13 | -0.42 | -0.36 | -1.16 | -1.03 | -0.89 | 0.05 | -0.05 | -0.31 | -1.10 | -1.01 | -0.97 |
| 운송 | TM-T | 0.05 | 0.11 | 0.16 | 0.26 | 0.27 | 0.50 | 0.48 | 0.43 | 0.81 | 0.38 | 0.42 | 0.65 | 1.70 | 1.83 | 4.16 |
| | BS-T | -0.03 | -0.03 | 0.33 | -0.12 | -0.02 | 0.08 | 0.04 | 0.14 | 0.23 | -0.57 | -0.11 | 0.25 | -0.62 | -0.04 | 1.21 |
| 은행 | TM-T | 0.12 | 0.16 | 0.23 | 0.33 | 0.69 | 0.91 | 0.35 | 0.35 | 0.10 | 0.46 | 0.84 | 0.88 | 1.92 | 3.83 | 3.85 |
| | BS-T | 0.28 | 0.38 | 0.42 | 0.32 | 0.53 | 0.56 | -0.47 | -0.14 | -0.02 | -0.11 | -0.23 | -0.18 | -0.21 | 0.41 | 0.78 |

Table 5. TM-T의 매매 성과 분석

| | 샤프 지수 | | | 젠센의 알파 | | | IR | | |
|---|---|---|---|---|---|---|---|---|---|
| | Δ1 | Δ2 | Δ3 | Δ1 | Δ2 | Δ3 | Δ1 | Δ2 | Δ3 |
| 건설 | 2.70 | 1.84 | 1.49 | 0.04 | 1.59 | 0.84 | 1.08 | 0.69 | 0.51 |
| 소비재 | 1.13 | 1.09 | 0.76 | 0.08 | 0.10 | 0.04 | 0.14 | 0.09 | -0.07 |
| 에너지 | 0.69 | 0.75 | 1.04 | 0.40 | 0.36 | 0.45 | -0.63 | -0.50 | 0.14 |
| 자동차 | 4.51 | -8.63 | -3.11 | 0.16 | 0.21 | 0.25 | 0.67 | 0.90 | 0.94 |
| 조선 | -0.14 | -0.22 | -0.18 | 0.10 | 0.45 | 0.14 | 1.21 | 1.37 | 1.23 |
| 철강 | -4.65 | -0.76 | -0.29 | 0.18 | 0.12 | 0.01 | 1.19 | 1.22 | 0.71 |
| 반도체 | 0.44 | 1.02 | 0.86 | 0.06 | 0.24 | 0.20 | -0.19 | 0.04 | -0.03 |
| IT | -1.98 | -1.31 | -1.37 | 0.25 | 0.25 | 0.36 | 1.27 | 1.33 | 1.40 |
| 미디어 | -0.51 | -0.45 | -0.72 | 0.07 | 0.30 | 0.69 | 1.22 | 1.49 | 1.61 |
| 운송 | -1.57 | -5.74 | 2.27 | 0.25 | 0.29 | 0.64 | 1.41 | 1.51 | 1.20 |
| 은행 | 2.91 | 3.53 | 2.62 | 0.30 | 0.50 | 0.49 | 1.01 | 0.98 | 0.84 |

측면에서 BS-T를 바탕으로 매매한 수익률에 비해서 우수하다. 또한 소비재와 철강의 섹터를 제외하고는 이동평균선기간이 클수록 누적 로그수익률이 높은데, 이는 일반적인 트렌드가 특정한 방향으로 움직이는 모멘텀을 발생하기 때문에 단기적인 트렌드 변화에 대응하여 매매하는 것 보다는 일정하게 유지되는 트렌드를 바탕으로 매매하는 것이 더 높은 수익률을 보인다는 것을 알 수 있다. 또한 단기적이고 민감한 트렌드 변화로 인한 빈번한 매매 타이밍은 매매 비용이 많이 발생한다. 그리고 동일한 섹터와 동일한 년도의 경우에는 Δ1 ~Δ3 간에 로그수익률에 대한 편차가 낮다. 예컨대 KODEX 건설의 경우, 2011년을 보면 $\Delta t$ 별로 TM-T의 로그수익률이 각각 약 –0.10, -0.12, -0.13이고, BS-T의 경우에는 $\Delta t$ 별로 각각 0.01, 0.44, 0.41이다. 반면 2012년도에는 TM-T의 경우에 $\Delta t$ 별 로그수익률이 모두 0.40 대 이다. 이처럼 동일한 섹터와 동일한 년도의 경우에 $\Delta t$ 별로 로그수익률 간의 편차가 작은 것을 알 수 있는데, 이는 주 별로 수집한 단어 별 포털 트렌드 값의 경우에 $\Delta t$ 간의 편차가 크지 않기 때문이다. 앞서 Table 1의 KOSPI 지수와 섹터 별 상관관계와 섹터 별 수익률 간의 유의한 상관관계는 없는 것으로 사료된다. 많은 경우의 수를 실험함으로써 각각에 대한 개별적 성과를 로그수익률을 통해서 확인할 수 있지만 종합적인 매매 성과를 정량적인 평가 척도로 이해하기는 쉽지 않다(신현준, 2013). 따라서 다음 절에서 일반적으로 포트폴리오의 운용 성과를 측정하는 샤프지수(Sharpe Ratio)와 젠센의 알파(Jensen's alpha) 그리고 정보비율(Information Ratio; 이하 IR)을 통해서 TM-T를 바탕으로 $\Delta t$ 별로 매매한 경우에 대해서 성과 평가를 진행한다.

### 5.3 성과 측정

금융 자산으로 구성된 포트폴리오의 운용 실적에 있어서 단지 운용 수익률로 운용에 대한 성과를 측정하는 것은 옳지 않다. 예컨대 시장 수익과 같이 벤치마크 수익 대비하여 좋은 성과를 보인다면 음의 수익이 발생하더라도 저조한 성과라고만 볼 수 없다. 또한 운용 수익률의 변동성, 매매 수수료의 형평성 그리고 포트폴리오 구성과 수정에 대한 방법론 등 다양한 측면으로 포트폴리오 운용 성과를 분석해야 한다. 일반적으로 정량적인 방법으로 포트폴리오 성과 측정을 위한 도구로 샤프지수와 젠센의 알파 그리고 IR이 있다. 샤프지수는 포트폴리오의 위험 1 단위에 대한 초과 수익의 정도를 나타내는 지표, 즉 초과 수익이 얼마인가를 측정하는 지표이고 모형은 식(1)와 같다.

$$\text{Sharpe Ratio} = \frac{R_i - R_f}{\sigma_i} \quad (1)$$

Notations :

$R_i$ 포트폴리오 $i$의 수익률
$R_f$ 무위험 수익률(국고채 3년 만기)
$\sigma_f$ 포트폴리오 $i$의 표준편차

젠센의 알파는 포트폴리오의 수익률이 균형 상태에서의 수익률보다 얼마나 높은지를 나타내는 지표, 즉 포트폴리오 수익률에서 기대 수익률을 뺀 값을 의미하며 모형은 식(2)과 같다.

$$\text{젠센의 알파} = (R_i - R_f) - b_p^*(K_i - R_f) \quad (2)$$

Notations :
$b_p$ : 포트폴리오의 베타
$K_i$ : 시장 수익률

그리고 IR은 포트폴리오 관리자의 능력을 측정할 수 있는 지표로 포트폴리오의 초과 수익률을 추적 오차로 나눈 값을 말하며 RVR(Reward-to-Variability Ratio)라고도 부른다. 세 가지 모두 측정 된 결과 값이 클수록 투자 성과가 우수하다고 할 수 있으며 IR의 경우에는 실무적으로 미국에서는 약 0.4~0.5 이내인 경우에 '우수'한 것으로 평가한다. IR의 산출 모형은 식(3)과 같다.

$$IR = \frac{(R_i - K_i)}{Te} \quad (3)$$

Notations :

$Te$ : 추적 오차의 표준편차

Table 5는 TM-T를 바탕으로 매매한 결과에 대한 성과를 측정한 결과를 $\Delta t$ 별로 보여준다. 특히 수익률이 높다고 성과 평가의 결과가 반드시 좋다는 것은 아니다. 예컨대 KODEX 건설의 경우에 $\Delta 3$ 의 경우가 가장 높은 수익률을 보이지만 샤프지수는 $\Delta 1$ , 젠센의 알파는 $\Delta 2$ 그리고 IR은 $\Delta 1$ 이 높게 나타난다. 이는 수익률을 통한 평가와는 다르게 성과 평가는 연도별 수익률의 표준편차와 벤치마크 대비 초과수익 등을 고려하기 때문에 반드시 누적수익률이 높다고 성과 평가 결과가 우수한 것은 아니다. 즉 우수한 운용 능력은 수익률과 함께 운용 위험(Risk)을 항상 고려해야 한다. 여기서 벤치마크는 BS-T로 설정하였다. 샤프지수 중에서는 약 84%의 누적수익률을 기록한 KODEX 자동차가 4.51($\Delta 1$ )로 가장 우수했고, 젠센의 알파는 약 400%의 누적수익률을 보인 KODEX 건설이 1.59($\Delta 2$ )로 가장 운용 성과가 높았다. 마지막으로 IR은 약 230%의 누적수익률을 보인 TIGER 미디어가 1.61($\Delta 3$ )로 가장 높았다.

## 6. 결 론

본 연구는 대중들의 감정이 개인행동과 의사결정에 큰 영향을 미칠 수 있다는 행동경제학의 이론을 토대로 국내 빅데이터 트렌드를 이용한 ETF 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 이론을 수립하였고, 국내 주식시장의 주가 및 네이버 트렌드(Naver trends) 검색량 데이터 변화량의 연관성을 이용한 ETF 투자전략을 제안하고 결과를 분석하였다. 또한 텍스트마이닝(Text Mining) 기법을 이용하여 섹터 별로 의미 있는 키워드(Keywaod)를 선정하였다. 이는 브레인스토밍 기법(Brain Storming Method)인 BS 기법을 통해서 선정된 키워드를 바탕으로 2011년부터 2014년까지 ETF를 매매한 수익률과 비교하였고, 그 결과 대체적으로 본 연구에서 제안한 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매했을 때 보다 더 높은 수익률을 보였다. 또한 시장 참여자들의 정보검색 움직임을 정량화하기 위해 검색 상대변화량 $\Delta 1 \sim \Delta 3$ 로 나눠 실험한 결과, 누적 로그수익률 측면에서는 $\Delta 3$ 가 다소 높은 성과를 보였다. 그러나 매매 운용에 따른 수익률만을 갖고 운용 성과를 정량적으로 판단하기는 한계가 있기 때문에 샤프 지수, 젠센의 알파, IR을 산출하고 $\Delta t$ 간의 성과를 분석하였다. 그 결과 샤프 지수, 젠센의 알파, IR 측면에서는 수익률이 다소 높았던 $\Delta 3$ 보다는 $\Delta 1$ 와 $\Delta 2$ 가 우수한 성과를 보였다. 이는 이동평균선 기간이 클수록 단기적인 사건으로 인한 트렌드 변화가 아닌 일정기간 유지되면서 포착되는 트렌드를 바탕으로 매매하는 것이 수익률 측면에서 더 우수하다는 것을 의미한다.

섹터 별로 선정된 키워드에 따라서 운용 성과가 다르게 나타난 본 연구의 결과를 통해서 투자자의 심리가 반영된 포털 트렌드를 바탕으로 매매 전략을 수립할 시, 투자 자산과 연관된 키워드를 효과적으로 선정하는 것이 무엇보다 중요하다는 것을 알 수 있었다.

## References


[1] 김유신, 김남규, 정승렬, “뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형”, 지능정보연구, 제18권, 제2호, pp. 143-156, 2012.

[2] 김정래, 정찬기, “전화통화 빅데이터 분석에 관한 연구”, 한국정보기술아키텍처논문지, 제10권, 제3호, pp. 387-397, 2013.

[3] 박원준, “‘빅데이터(Big Data)’ 활용에 대한 기대와 우려”, Journal of Communications & Radio Spectrum, 제51권, pp. 28-47, 2012.

[4] 신현준, 라현우, “금융시장의 빅데이터 트렌드를 이용한 주가지수 투자 전략”, 한국경영과학회지, 제32권, pp. 91-103, 2015.

[5] 신현준, 유재필, “DEA-마코위츠 결합 모형을 이용한 건설기업의 효율적 포트폴리오 구성 방안”, 한국산학기술학회, 제14권, pp. 899-904, 2013.

[6] 이득환, 김수현, 강형구, “빅데이터를 사용한 시스템 트레이딩: KOSPI200 선물을 대상으로”, 한국재무학회, 2014.

[7] 임유진, 최은미, “시계열 빅데이터 처리 분석을 위한 맵리듀스 메커니즘 연구”, 한국정보기술아키텍처논문지, 제12권, 제1호, pp. 97-98, 2015.

[8] 옥기율, 김지수, “소비자 심리지수가 KOSPI 수익률에 미치는 비대칭적 영향에 대한 연구”, 금융공학연구, 제11권, 제1호, pp. 17-37, 2012.

[9] Barber, B. M., Odean, T. and Zhu, N., “Systematic Noise”, Journal of Financial Markets, Vol. 12, pp. 547-569, 2009.

[10] Bordino, I., Battiston, S., Caldarelli, G., Cristell, M. and Ukkonen, A., “Web Search Queries Can Predict Stock Market Volumes”, PloS ONE, Vol. 7, No. 7, pp. 1-17, 2012.

[11] Choi, H. and Varian, H., “Predicting the Present with Google Trends”, The Economic Record, Vol. 88, pp. 2-9, 2012.

[12] Dorn, D., Huberman, G. and Sengmueller, P., “Correlated Trading and Returns”, Journal of Finance, Vol. 63, pp. 885-920, 2008.

[13] Jackson, A., “The Aggregate Behaviour of Individual Investors”, London Business School, 2003.

[14] Kumar, A. and Lee, C. M. C., “Retail Investor Sentiment and Return Comovements”, Journal of Finance, Vol. 61, No. 5, pp. 2451-2486, 2006.

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

[16] Preis, T., Moat, H. S. and Stanley, H. E., “Quantifying Trading Behavior in Financial Markets Using Google Trends”, Scientific Reports, Vol. 3, 2013.

**유재필(Jae Pil Ryu)**
상명대학교 학사, 석사
현재: KIS채권평가 선임연구원, 상명대학교 박사과정
관심분야: 금융공학, 데이터마이닝
E-mail: jaepilryu@kispricing.com

**한창훈(Chang Hoon Hahn)**
경북대학교 전자공학과
서강대학교 소프트웨어공학과
현재: 상명대학교 박사과정
관심분야: 금융공학, 데이터마이닝
E-mail: hahn65@gmail.com

**신현준(Hyun Joon Shin)**
고려대학교 학사, 석사, 박사
미국 Texas A&M 대학교 박사후연구원
(주)삼성전자 책임연구원
현재: 상명대학교 경영공학과 부교수
관심분야: 금융공학, 조합최적화 응용, 데이터마이닝
E-mail: hjshin@smu.ac.kr