한국경영과학회지
제38권 제4호
2013년 12월
http://dx.doi.org/10.7737/JKORMS.2013.38.4.053

인터넷 검색추세를 활용한 빅데이터 기반의 주식투자전략에 대한 연구*

김민수 · 구평회†
부경대학교 시스템경영공학부

A Study on Big Data Based Investment Strategy Using Internet Search Trends

Minsoo Kim · Pyunghoi Koo
Department of Systems Management and Engineering, Pukyong National University

■ Abstract ■

Together with soaring interest on Big Data, now there are vigorous reports that unearth various social values lying underneath those data from a number of application areas. Among those reports many are using such data as Internet search histories from Google site, social relationships from Facebook, and transactional or locational traces collected from various ubiquitous devices. Many of those researches, however, are conducted based on the data sets that are accumulated over the North American and European areas, which means that direct interpretation and application of social values exhibited by those researches to the other areas like Korea can be a disturbing task.

This research has started from a validation study against Korean environment of the former paper which says an investment strategy that exploits up and down of Google search volume on a carefully selected set of terms shows high market performance. A huge difference between North American and Korean environment can be eye witnessed via the distinction in profit rates that are exhibited by the corresponding set of search terms. Two sets of search terms actually presented low correlation in their profit rates over two financial markets. Even in an experiment which compares the profit rates with two different investment periods with the same set of search terms showed no such meaningful result that outperforms the market average. With all these results, we cautiously conclude that establishing an investment strategy that exploits Internet search volume over a specified word set needs more conscious approach.

Keyword : Big Data, Stock Investment Strategy, Google Trends, Naver Trends, Search Terms, KOSPI

논문접수일 : 2013년 10월 01일 논문게재확정일 : 2013년 12월 09일
논문수정일(1차 : 2013년 11월 30일)
* 이 논문은 부경대학교 자율창의학술연구비(2013년)에 의하여 연구되었음.
† 교신저자 phkoo@pknu.ac.kr

1. 서 론

빅데이터에 대한 명확한 정의는 연구자나 관련 기업들에 따라 차이를 보이지만, 일반적으로 기존 데이터에 비해 그 크기가 너무 커서 통상적인 방법으로는 수집하거나 분석하기가 어려운 데이터 집합체를 가리키는데 사용된다[1]. 스마트 폰과 센서 장비들의 증가로 인해 생성되는 데이터의 종류와 규모가 폭발적으로 증가함에 따라 이러한 데이터의 단순한 축적에서 보다 나아가 이를 다양한 사회적 가치 특히 비즈니스적인 가치에 활용하고자 하는 시도가 빅데이터에 대한 높은 관심의 배경이라고 할 수 있겠다.

빅데이터와 관련되어 발표되는 연구 결과 중에서 많은 내용들이 검색 엔진이나 블로그, SNS(Social Network Service) 등을 통해 입력되는 단어 집합들의 추세적인 흐름을 활용하여 관련 대상들의 사회적 혹은 경제적 경향을 예측하는 것과 관련되어 있다. 여기에는 인터넷 검색 서비스를 사용한 독감 예보에 대한 연구[11, 15], 년도 별 검색 성향과 해당 국가의 GDP와의 상관관계를 살펴 보는 연구[17], 특정 단어집합에 대한 검색률의 증감이나 Wikipedia 사용패턴을 통해 금융 시장에 대한 참가자들의 행태를 파악하고자 하는 연구[13, 16, 18], 검색 횟수와 선거 결과의 연관성을 찾는 시도[3] 등 매우 다양한 아이디어들이 실험되어 발표되고 있다. 일찍부터 검색결과에 대한 조회 서비스를 시작하여 상대적으로 긴 기간 동안의 검색 결과를 제공하는 외국의 상황과는 달리, 국내 포탈에서 검색결과에 대한 추세조회 서비스를 제공하기 시작한 것은 최근의 일이며, 검색 가능 기간 또한 짧아서 국내의 빅데이터와 관련된 연구들은 비교적 최근에 진행되기 시작했다.

아직까지 많은 연구들이 주로 해외의 연구를 통해 발표되는 상황에서, 이들 연구 결과를 그대로 받아들이기에는 해외와 국내의 상황이 많은 차이를 보인다는 것을 미리 이해할 필요가 있겠다. 일반적으로 인터넷 검색어나 사회망에 따른 관계의 분석을 통해 드러나는 2개 이상의 상이한 요소들의 연관성은 객관적인 논리에 기초해서 설명되기 보다는 주어진 데이터 집합 내에서 나타나는 현상을 개연성에 기초하여 설명하는 형태이기 때문에, 각기 다른 상황과 가정 하에서 축적된 데이터 집합에서 이러한 연관성을 발견하고자 하는 시도는 전혀 다른 양상을 보이거나 심하게는 유의미한 연관성을 찾을 수 없는 결과를 보게 된다.

본 연구는 Preis et al.[18]의 연구 결과를 국내 상황의 특수성과 함께 살펴보고자 하는 시도에서 시작되었다. Preis et al.은 구글트렌드(Google Trends)를 이용하여 용어 별 검색수의 추이가 주식시장의 향후 움직임을 예측하는데 활용될 수 있다는 가정 하에서, 98개의 단어를 선별하여 이 단어집합의 검색수 증감을 기초로 한 투자전략을 세우고, 이 전략이 시장평균 이상의 수익률을 보인다는 결과를 제시하였다. 그러나 이 연구는 미국의 데이터 집합을 중심으로 다우존스산업지수(DJIA : Dow Jones Industrial Average)를 대상으로 이루어졌으며, 무엇보다도 기준이 되는 검색어의 선정에 있어서 범국가적인 특성을 반영한 것이 아니라는 점에서 해당 연구 결과를 활용하는데 있어서 객관적인 재검토가 필요하다. 본 연구에서는 이러한 기존 연구의 결과가 한국의 주식 시장에서 직접적으로 활용되기 어렵다는 점을 살펴보고, 한국적인 경제 상황을 반영할 수 있는 검색어 집합의 선택에는 보다 많은 연구가 필요함을 얘기하고자 한다. 더 나아가서는 과연 검색어에 기초한 시장 투자 전략이 효과적인 방법이 될 수 있는 지에 대한 토론의 화두를 제기한다는 점에서 의미를 가질 수 있겠다.

본 논문의 구성은 다음과 같다. 먼저 제 2장에서는 빅데이터의 개념을 간략히 소개하고, 빅데이터와 관련된 연구 중에서 연관성 파악을 다루는 연구를 중심으로 기존 연구 내용을 정리하였다. 제 3장에서는 Preis et al.[18]의 기존 연구를 한국 시장 환경에 대응시켜 재검토하는 실험 분석 내용을 다룬다. 마지막으로 제 4장에서는 본 연구의 결론과 함께 실험결과를 토대로 추가적인 논의 사항을 다루었다.

2. 관련 연구 및 배경

2.1 빅데이터의 개념

스즈키 료스케는 빅데이터를 '사업에 도움이 되는 인사이트를 도출하기 위해 고해상, 고빈도로 생성되는 다양한 데이터'라고 정의하면서, 빅데이터의 3가지 특징을 고해상, 다양성 및 고빈도로 정의하였다[1]. 기존에는 한데 묶어서 다루어왔던 현상을 각각의 요소로 분해하여 파악하고 대응할 수 있는 데이터라는 의미에서 고해상이란 특징을 부여하였고, 취득이나 생성 혹은 처리 대상이 되는 데이터의 사이즈가 크지는 않더라도 매우 높은 빈도로 생성된다는 의미에서 고빈도 생성의 특징을 부여하였다. 마지막으로 정형적인 수치 데이터, 텍스트 데이터에 그치지 않고 웹 서비스를 이용한 유저의 기록, 방범 카메라 영상, 디지털 사이니지(Signage)를 보는 사람의 얼굴 사진, 위치정보, 각종 센서에서 수집된 데이터 등 다양한 종류의 데이터를 연계하여 활용하게 된다는 측면에서 다양성을 제시한 것이다. 빅데이터의 개념은 처음에는 데이터 규모와 기술적 측면에서 시작되었지만 차츰 그 가치와 활용효과 측면에서 관찰되면서 그 개념도 확대되는 추세에 있다[2]. 이와 유사하게 한국정보화진흥원에서도 가트너의 보고서를 인용하며, 빅데이터의 4대 특징으로 규모, 다양성, 생성 속도 및 복잡성을 언급하기도 하였다[4].

빅데이터와 관련된 연구에 대해서 단순히 비즈니스 인텔리전스나 데이터 마이닝의 새로운 마케팅 용어에 불과하다는 인식이 존재하는 것도 사실이지만, 주로 정형화된 데이터를 중심으로 비즈니스적 가치의 발견에 초점을 두어 진행되는 기존의 접근과는 달리, 비정형의 다양한 정보 소스로부터 취합되는 비사업적 성격의 정보를 다양한 사회적 가치의 발견에 초점을 두어 다룬다는 점에서 차별화되는 성격이 분명히 존재한다고 하겠다. 물론 사용되는 기법의 측면에서 볼 때, 데이터 마이닝의 여러 기법에 크게 의존하고는 있지만, 통계적 혹은 기계학습에 기반한 알고리즘적인 접근뿐만 아니라 상황인지적인 연결관계 속에서 정보를 해석하려는 다양한 접근법을 함께 고려하고 있다는 점에서 보다 포괄적인 접근법을 사용하고 있다고 하겠다.

2.2 인터넷 검색과의 연관성 발견에 대한 빅데이터 연구

현대인이 자신의 생활과 건강 등에 어떤 기대와 걱정이 있는 경우 또는 가까운 미래에 어떤 사항에 대하여 의사결정을 할 경우 대부분은 우선 인터넷 웹사이트에서 관련된 단어를 검색하는 경향이 있다. 예를 들어 여행을 준비 중인 사람이라면 생각하고 있는 여행지의 관광명소나 호텔을 인터넷에서 검색하게 될 것이다. 이와 같은 맥락에서 생각해 본다면, 어떤 여행지의 관광명소나 호텔에 대한 인터넷 검색 수가 증가하면 가까운 미래에 이곳의 여행객이 증가할 수 있다는 예측이 가능하다. 이러한 개념을 바탕으로 최근 다양한 분야에서 웹 검색 정보를 활용한 연구가 시도되고 있다.

가장 초기에 Ettredge et al.[10]는 실업과 관련된 인터넷 검색 수와 실제 미국 정부가 발표한 실업률과의 연관관계를 연구하였다. Choi and Varian[7]은 구글트렌드 데이터가 실업수당을 받기 위해 최초로 신청하는 사람들에 대한 사전 지표로 활용될 수 있음을 보였다. 의료건강 분야에서도 인터넷 검색 데이터를 활용한 연구가 활발히 연구되고 있다. Cooper et al.[9]은 여러 종류의 암에 대한 인터넷 검색 수와 실제 암 환자 발생 수와의 관계를 조사하였다. 암 관련 용어의 인터넷 검색 수는 암 발생건수와 상관성(상관계수 0.5)이 있으나 그 보다는 신문지 상에서 암과 관련된 뉴스의 빈도(상관계수 0.88)와 더 큰 연관성이 있다는 것을 발견하였다. 이를 통하여 검색빈도가 실제 암 발생에 대한 직접적인 연관성보다 다른 요인(신문 뉴스 빈도)에 의해 영향 받을 수 있으므로 관련성 분석에 있어 여러 가지 면을 고려해야 한다는 점을 강조하였다. Polgreen et al.[15]과 Ginsberg et al.[11]은 각각 Yahoo와 Google

에서 독감(Influenza)과 연관된 검색 결과와 실제 독감발생건수와의 관계를 조사하고, 웹 검색데이터를 이용하면 보건당국보다 먼저 독감의 유행을 예측할 수 있다는 것을 사례를 통하여 주장하였다. 실제로 Google에서는 이러한 개념을 이용하여 현재 Google Flu Trends(http://www.google.org/flutrends/)를 통하여 국가별로 독감유행정보를 실시간으로 일반인에게 공개하고 있다. 이외에도 검색엔진에서의 검색용어를 활용하여 노동과 주택시장[13], 영화, 게임 및 음악산업[12], 자동차 및 주택[8] 등 다양한 산업에 활용하려는 연구들이 시도되고 있다.

많은 적용 분야들 중에서도 특히 주식금융산업에 웹 검색 결과를 활용하는 방안에 대한 연구가 최근 활발히 진행되고 있다. Preis et al.[16]은 S&P500 기업의 주간 주식거래량이 대응되는 회사의 Google을 통한 인터넷 검색횟수와 연관관계가 있음을 발표하였다. 유사한 연구로서 Bordino et al.[6]은 NASDAQ에 상장되어 있는 기업의 주식거래량과 해당 기업과 관련된 용어가 Yahoo 검색엔진에서 검색된 횟수와의 관계를 비교분석하고, 이 둘간에는 상호 연관성이 존재함을 보였다. Bollen et al.[5]은 twitter.com에서의 감성(mood)과 관련된 용어의 빈도가 DJIA에 영향을 미친다는 것을 발견하고, DJIA를 예측하는 데에 twitter의 감성관련 단어를 고려한다면 정확도를 높일 수 있다고 주장하였다. Moat et al.[14]은 주식시장의 움직임이 있기 전에 인터넷 백과사전인 Wikipedia에서 관련된 용어에 대한 검색 빈도가 어떻게 변하는지에 대한 연구를 수행하였다. Preis et al.[18]은 특정 용어에 대하여 현재의 인터넷 검색 횟수가 가까운 미래의 주식시장의 가격등락과 관계가 있는지 확인하는 연구를 수행하였다. 우선 주식시장과 관련된 98개의 용어를 선정하고, 각 용어에 대하여 구글트렌드로 검색 추이를 얻은 후, 과거 일정기간보다 검색 횟수가 증가하면 DJIA 매도포지션을 취하고 반대면 매수포지션을 취한 후 일 주일 지나서 청산하는 방식으로 투자하는 전략을 제시하였다. 연구결과 구글트렌드의 데이터는 주식시장의 현재뿐만 아니라 가까운 미래의 상황도 예측하는데 도움이 된다는 것을 발견하고, 특히 주식시장이 폭락하기 전에 주요용어의 검색수가 증가한다는 것을 발견하였다.

위와 같이 발표된 다양한 연관성에 대한 연구 중에서, 본 논문에서는 Preis et al.[18]이 제시한 트렌드 투자전략이 국내에서도 적용될 수 있는지를 검증하는 것에 목적을 두고 있다. 동일한 투자전략이 두 주식시장에서 보이는 결과를 비교한 다음, 차이가 나타나는 결과에 대한 해석을 통해 국내 주식시장의 특징을 반영한 새로운 검색트렌드 기반의 투자전략이 어떻게 구성될 수 있을지에 대해서 논의한다.

3. 실험 분석

국내에서 인터넷 검색용어의 추이와 주가지수의 관계를 알아보기 위하여 단어 별 검색 추이는 검색하는 단어의 역사적 추이 정보를 제공해 주는 네이버트렌드(http://trend.naver.com/)를 이용하였고, 주가지수는 KOSPI200 지수를 사용하였다. 네이버트렌드는 2007년 이후 주단위로 단어 별로 검색횟수에 대한 추이정보를 제공하므로 KOSPI200도 주단위의 주가정보를 사용한다. 실시간으로 제공되는 인터넷 통계데이터(http://trend.logger.co.kr/)에 의하면 2012년 한해 동안 국내의 검색사이트의 검색 점유율은 naver.com이 74.5%를 차지하고 있다. 따라서 본 논문에서는 신뢰도 측면에서 검색수가 가장 많은 네이버트렌드의 추이데이터를 택하여 분석을 실시하였다. 실험의 주요 목적은 아래와 같다.

1. 국내 데이터를 이용한 기존 연구 결과의 검증 분석 : Preis et al.[18]는 구글트렌드를 이용하여 미국에서 용어 별 검색수의 추이를 이용하여 주식시장의 미래를 예측할 수 있다고 주장하였다. 본 실험은 이러한 내용이 국내에서도 활용될 수 있는지 동일한 단어를 가지고 비교 분석한다.
2. 검증데이터를 활용한 용어 별 투자수익률 검증 : 네이버트렌드를 이용한 주식거래전략을 위하여 초기 5년간(2007년~2011년)의 데이터를 학습데

이터로 하여 수익률과 이후 1년 7개월(2012년~2013년 7월 말) 동안의 수익률 간의 상관관계를 분석하여 제시된 용어를 기반으로 하는 트렌드 정보를 가지고 투자전략을 세웠을 때 원하는 결과가 나오는지 분석한다.

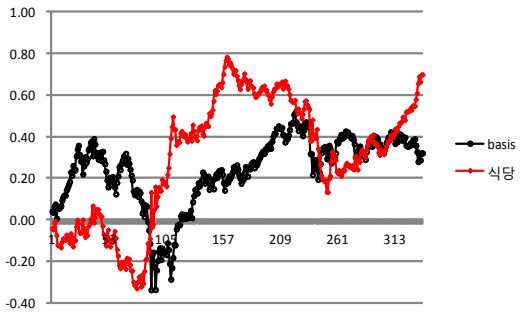
3.1 국내 데이터를 이용한 기존 연구 결과의 검증 분석

Preis et al.[18]의 연구에서는 98개의 단어를 선정하여 각 단어별로 구글트렌드에서 검색수의 변화가 주식시장에서의 의사결정에 영향을 주는지에 대한 실험을 수행하였다. 검색용어 트렌드를 기반으로 하는 투자전략은 다음과 같다.

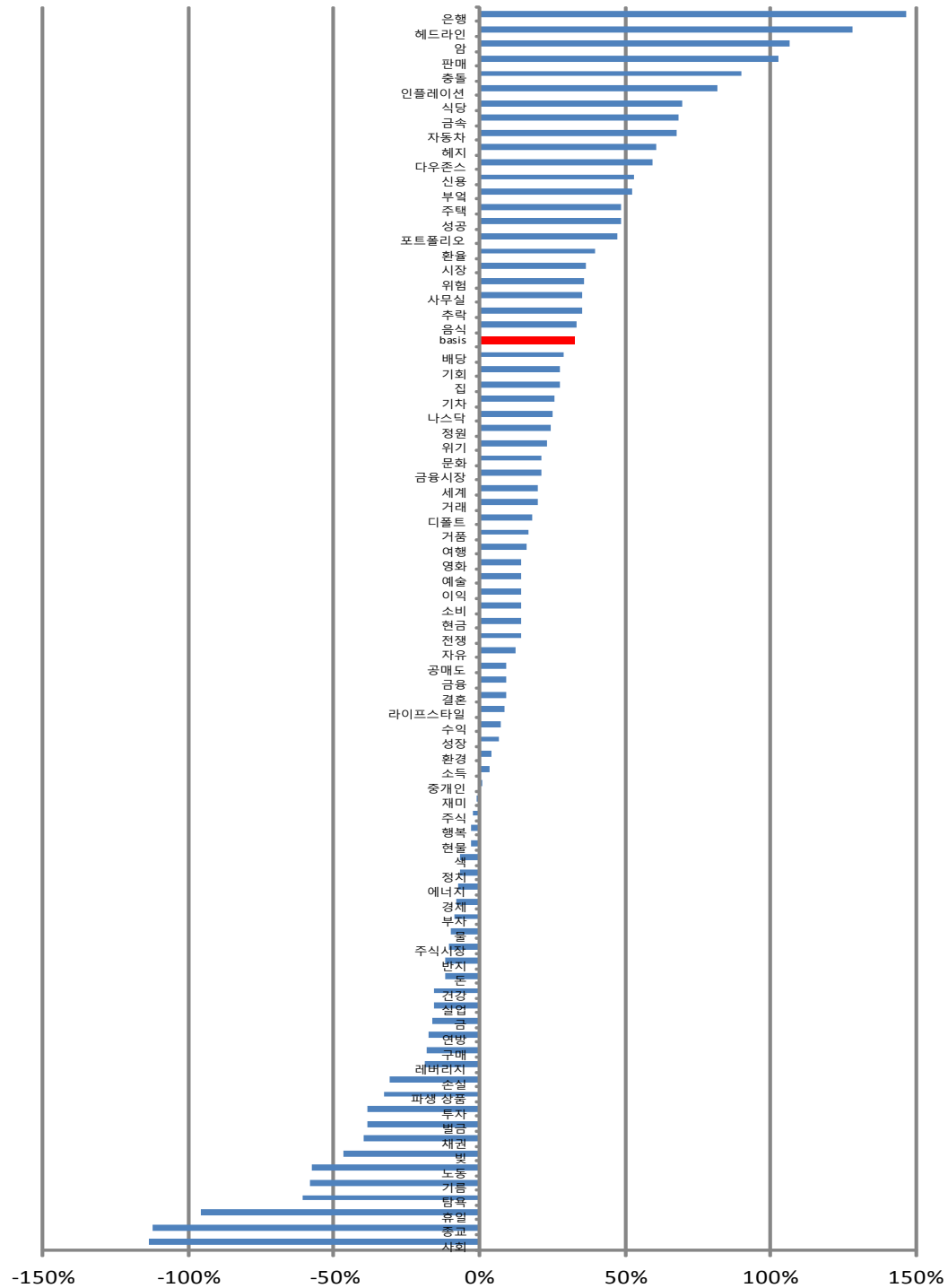
투자 의사결정을 해야 하는 현시점을 t주초라 하자. 구글트렌드에서는 특정한 단어의 검색 수에 대하여 주간별로 검색 추이를 보여준다(네이버트렌드에서도 주간단위로 상대적인 단어검색 추이 정보를 제공한다). (t-1)주간에 검색된 특정한 단어에 대한 상대적인 검색수를 n(t-1)이라고 하자. 그 이전의 Δt주, 즉 (t-1-Δt)주부터 (t-2)주까지의 평균 검색수를 계산하고 이를 N(t-2, Δt)로 표현한다. 즉, N(t-2, Δt) = {n(t-2)+n(t-3)+ … +n(t-1-Δt)}/Δt에 의해서 구한다. 이제, n(t-1)과 N(t-2, Δt)를 비교하여 n(t-1) > N(t-2, Δt), 즉 t-1주 한 주간의 검색수가 그 이전 Δt 동안의 평균 검색수보다 증가했다면 t주의 첫 번째 거래일 종가 p(t)에 다우존스 산업지수(DJIA)를 매도하고 다음 주 첫 거래일의 종가 p(t+1)에 DJIA를 환매(repurchase)하여 거래를 소멸시키는 전략을 취한다. 반대로 이전 Δt주 동안의 평균 검색수와 비교하여 검색수가 감소했다면 t주의 첫 번째 거래일 종가에 DJIA를 매수하고 다음 주 첫 거래일의 종가에 DJIA를 전매(resale)하여 거래를 소멸시킨다. 이러한 투자전략을 매주 초에 반복적으로 시행한다.

주간 수익률의 계산은 t주에 매도포지션을 취하는 경우 p(t)에 매도하고 p(t+1)에 매수하므로 기간 수익률은 p(t)/p(t+1)-1에 의해 계산되지만 Preis et al.[18]의 논문에서와 마찬가지로 매도포지션이나 매수포지션이 누적수익률에 동일하게 영향을 주도록 자연 log 함수를 사용하여 log(p(t))/log(p(t+1))를 사용한다. 매수포지션에서의 수익률은 따라서 log(p(t+1))/log(p(t))을 사용한다. 이처럼 주간수익률을 계산하면 누적수익률은 주간수익률을 더하여 간단히 구할 수가 있다. 국내와 미국의 결과를 비교하기 위하여 98개의 단어를 한국어로 번역하고 번역시 한국어에서는 거의 쓰이지 않거나 두 영어단어가 하나의 한국어로 표현될 때는 단어를 제거하여 분석대상 단어는 총 84개로 하였다.

[그림 1]은 '식당(restaurant)'이란 검색용어를 이용한 투자전략을 사용한 경우 2007년 1월 초(1주)에서 2013년 7월 말(343주) 까지 6년 7개월 간의 누적수익률을 보여주고 있다. 여기서 벤치마크로 사용하는 basis는 buy-and-hold 전략으로 주초 첫 거래일 종가로 항상 주식을 매수하고 그 다음 주 첫 거래일 종가로 매도 청산하는 전략을 사용하는 경우의 누적수익률을 나타내고 있다. 이 예에서는 벤치마크인 basis가 32.1%의 수익률을 보인 반면에 '식당'을 가지고 네이버트렌드 투자전략을 사용한 경우에는 69.9%의 수익률을 보이고 있는 것을 알 수 있다.



[그림 1] 검색 용어 '식당'을 이용한 네이버트렌드 투자전략의 수익률 추이 비교(x축은 주단위 시간을 나타내고, y축은 해당 주까지의 누적수익률을 나타냄)



[그림 2] 검색용어별 트렌드 투자전략의 누적수익률(x축은 2007년 1월 첫 주에서 2013년 7월 말까지의 누적수익률을 나타냄)

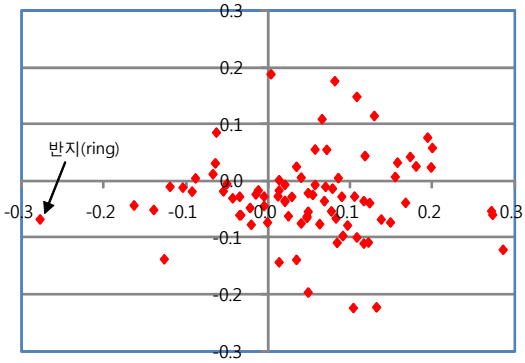
[그림 2]는 84개의 검색용어에 대한 총 대상기간의 누적수익률을 나타내고 있다. 검색 용어 중 '은행'이 가장 높은 수익률인 146.5%의 수익률을 보여주었고, '헤드라인', '암', '판매'등이 100% 이상의 누적수익률을 실현하였다. 반면에 검색용어 '사회'를 사용한 트렌드 투자전략은 -113.4%의 누적수익률을 보이고 있고, '종교'도 -100% 이하의 누적수익률을 보이고 있다. 반면에 벤치마크로 사용하고 있는 buy-and-hold 투자전략의 basis는 32.1%의 누적수익률로 23번째로 높은 수익을 보이고 있다.

앞의 [그림 2]에서의 네이버트렌드 투자전략에서 어떤 용어를 사용하는가에 따라 수익률이 달라진다는 것을 볼 수 있다. 이러한 수익률을 각 검색용어별로 Preis et al.[18]의 결과와 비교하여 조사해 보면 두 나라 사이에 많은 차이가 남을 알 수 있다. 예를 들어 검색용어 '종교(religion)'의 경우 구글트렌드를 활용한 투자전략에서는 높은 수익률을 보인 반면에 네이버트렌드를 이용한 국내 수익률은 대규모 마이너스 수익률을 기록하고 있다. 구글트렌드 투자전략에서 가장 높은 수익률을 보인 'debt(빚)'은 네이버트렌드에서는 -46.7%의 좋지 않은 수익률을 기록하고 있다. 이처럼 구글트렌드와 네이버트렌드의 수익률 차이를 분석하기 위하여 본 논문에서는 각 검색용어에 대한 두 나라에서의 수익률 간의 관계를 조사하였다.

우선, 동일한 단어에 대한 미국에서의 수익률과 국내에서의 수익률의 관계성을 공정하게 분석하기 위해 동일한 기간인 1년간의 평균 수익률을 가지고 비교하였다. 또한 시장자체의 변화에 대한 영향을 없애기 위하여 벤치마크 수익률을 감안하여 다음과 같이 조정연수익률을 산출하여 비교하였다 :

조정연수익률 = (트렌드 투자전략 누적수익률
-벤치마크 누적수익률)/대상기간(년).

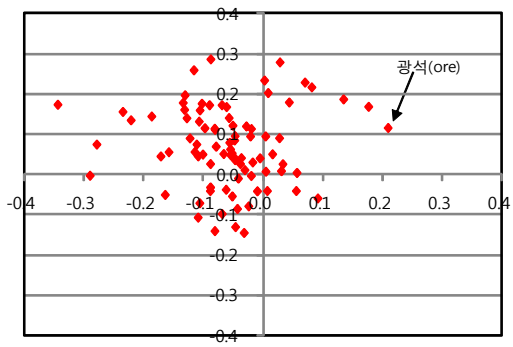
[그림 3]은 선택된 84개의 단어에 대하여 구글트렌드 투자전략을 적용한 미국 DJIA 조정연수익률(*x*축)과 네이버트렌드를 적용한 국내 KOSPI 지수 조정연수익률(*y*축)의 관계를 보여주는 산점도이다. 각 점들은 하나의 특정 검색용어에 대한 수익률이다. 예를 들어 가장 왼쪽에 위치한 점은 검색용어 '반지(ring)'를 이용한 구글트렌드 투자전략의 결과로서, 미국 DJIA는 -27.8%의 조정연수익률 보인데 반하여 국내 KOSPI는 -6.8%의 조정연수익률을 나타내고 있다. 그림에서 볼 수 있듯이 기대와는 다르게 주어진 단어들에 대한 미국과 국내에서의 연관성은 상당히 낮다는 것을 알 수 있다. 실제로 두 국가간의 검색단어를 이용한 트렌드 투자전략 수익률의 상관계수는 0.02로 관계가 미미하였다. 즉 Google을 이용한 미국에서의 트렌드 투자전략을 활용한 주식거래와 동일한 단어를 국내에서 직접 적용하는 것은 유용한 투자전략이라고 볼 수 없다는 것을 의미한다. 본 논문의 저자들은 이러한 실험 결과를 두 가지 요인에 기인할 수 있다고 판단하였다. 첫째는 검색 용어를 사용하여 투자하는 것이 미국에서와는 다르게 국내에서는 적절하지 않을 수 있다는 점과, 두 번째는 미국과 국내의 주식시장은 다른 패턴으로 움직이고 시장환경도 상이하므로 주식투자와 관련하여 투자자들이 관심 있어 하는 단어가 양국이 다를 수 있다는 것이다. 이러한 판단을 검증하기 위하여 다음과 같은 추가적인 실험을 수행하였다.



[그림 3] 트렌드 투자전략을 활용한 미국과 국내의 조정연수익률 관계를 나타낸 산점도(x축은 구글트렌드 투자전략을 이용한 DJIA의 검색용어별 연간조정수익률을 나타내고, y축은 네이버트렌드를 이용한 KOSPI200의 검색용어별 연간조정수익률을 나타냄)

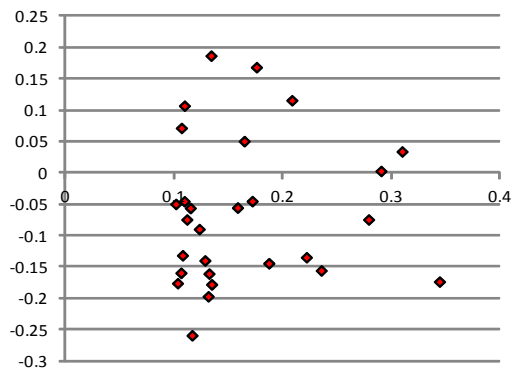
3.2 검증데이터를 활용한 검색용어 별 투자수익률 검증

우선 현재 고려하고 있는 검색용어를 대상으로 투자전략의 형성을 위하여 전체 데이터를 학습데이터 세트와 검증데이터세트로 구분한다. 즉, 2007년~2011년 사이의 데이터를 학습데이터로 하여 투자전략을 세우고, 이를 기반으로 투자하였을 경우에 2012년~2013년을 검증데이터로 사용하여 결과가 어떤지를 알아보는 실험을 수행하였다. [그림 4]는 이 두 기간 동안에 검색단어별로 조정연수익률을 나타내고 있다. 예를 들어 오른쪽 끝 점은 광석(ore)에 대한 연 수익률이 2007년~2011년 동안에는 20.8%이었고 2012년~2013년에는 11.6%라는 것을 보여주고 있다. 그래프를 통해 알 수 있듯이 모델 형성기간에 수익률이 높다고 하여 검증기간에도 수익률이 높다고 할 수 없다. 이 데이터의 상관계수는 -0.01으로 대단히 낮아 상관성이 없다고 할 수 있다.



[그림 4] Preis et al.[18] 검색단어별 연수익률 관계를 나타내는 산점도(x축은 학습데이터 세트(2007년~2011년)를 이용한 연간조정수익률을 나타내고, y축은 검증데이터 세트(2012년~2013년)를 이용한 연간조정수익률을 나타냄)

추가 실험에서 대상으로 하는 84개 검색용어 중에서 일정한 수익률을 내는 단어들을 선택하여 실험결과를 분석하였다. [그림 5]는 조정연수익률이 10% 이상 되는 단어만을 선택하여 트렌드 투자전략을 사용한 투자 결과이다. 모형 형성 단계에서 10% 이상의 손실을 본 단어들은 투자전략을 역으로 하여 (즉 매수포지션 전략) 투자하는 것으로 하였다. 대상 단어는 28개로 이들 단어들을 모두 포함하여 투자하는 경우 조정연수익률은 16.5%이었다. 그림에서 x축은 학습기간(2007년~2011년) 동안 트렌드 투자전략을 적용했을 경우의 조정연수익률이고 y축은 동일한 투자전략을 가지고 검증기간(2012년~2013년) 동안 투자한 결과이다. 그림에서 볼 수 있듯이 높은 수익률이 기대되는 검색용어만을 가지고 트렌드 투자전략을 사용하여 투자하는 경우에도 과거의 수익률과 기대되는 미래의 수익률이 관련성이 없다는 것을 볼 수 있다. 실험결과 검증기간의 평균적인 조정연수익률은 -6.3%로 투자손실을 보게 되는 결과를 얻었으며, 단어 별 조정수익률의 상관 계수도 0.06으로 상당히 낮은 것으로 나타났다. 이는 선택된 현재의 단어를 가지고 과거에 투자결과가 좋게 나왔다고 하여 미래에도 수익률이 좋을 것이라고 예측하는 것은 위험하다는 것을 말하고 있다.



[그림 5] 연 수익률 10% 이상을 내는 검색단어에 대해 학습기간(2007년~2011년)과 검증기간(2012년~2013년 7월)의 연수익률 관계를 나타내는 산점도(x축은 학습기간 동안의 연간조정수익률을 나타내고, y축은 검증기간 동안의 연간조정수익률을 나타냄)

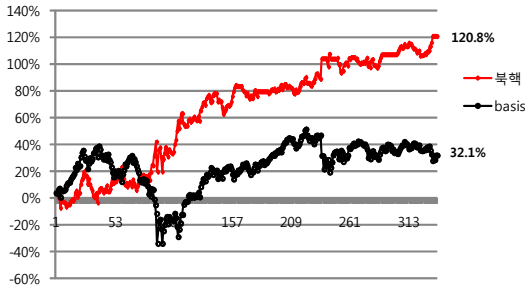
4. 결론 및 논의 사항

실험을 통하여 국내에서 네이버트렌드를 이용한 투자전략은 미국에서 구글트렌드를 이용한 투자전

략과는 상이한 결과를 냄을 알 수 있었다. 일부 검색 용어는 벤치마킹 수익률 보다 높은 수익률을 주지만 상관분석을 통한 분석결과 과거에 좋은 결과를 주었다고 하여 미래에도 동일하게 적용되지 않음을 실험을 통하여 확인할 수 있었다. 구글트렌드를 이용하여 시장평균 이상의 수익률을 보이는 검색어가 전체의 2/3가량을 차지하였지만, 네이버트렌드를 이용한 경우에는 그 수가 약 20개 정도로 전체 검색어 84개의 약 25% 정도에 불과하였다. 양쪽 모두에서 시장평균 이상의 수익률을 보이는 검색어는 약 8개 정도로 전체 대상 검색어의 10%에도 미치지 못하는 결과를 보였다.

인터넷 검색어를 활용한 투자전략이 외국과 국내에 대해서 서로 다른 결과를 보이는 원인에 대해서는 많은 추측이 가능하다. 먼저 두 언어의 단어가 가지는 외연의 크기가 서로 다르기 때문에, 외국의 연구에서 사용된 검색 단어에 정확히 대응되는 한국어 단어를 선택하는 것이 문제가 될 수 있다. 예를 들어, 영어의 Debt는 사전상의 의미로 빚, 부채, 채무 심지어는 신세짐, 은혜, 의리와 같은 다양한 의미로 한국어에 대응된다. 이처럼 단어들 간의 의미상의 차이로 인해 정확한 대응관계를 갖는 하나의 검색어를 찾는 것에는 근본적인 한계가 있으므로, 하나의 영어 검색어에 대응되는 여러 한국어 검색어들을 복합적으로 활용하여 그 결과를 비교해 보는 추가적인 실험을 고려해 볼 수 있겠다.

또 다른 원인으로 생각해 볼 수 있는 것은 비교 대상이 되는 미국과 한국의 경제적, 문화적 환경이 큰 차이를 보인다는 점이다. 국내 주식시장은 다른 나라와는 다른 독특한 특성을 가지고 있다. 남북관계, 수출주도형산업, 상대적으로 높은 제조업산업 비중, 심한 외국의존도, 환율변동위험 등은 우리 경제(따라서 주식시장)에 상대적으로 커다란 영향을 미친다. 이러한 국내 금융시장의 특이성을 고려하여 검색트렌드 투자전략을 사용하면 의미 있는 결과가 나올 수 있을 것이라 기대된다. 예를 들어, [그림 6]은 남북관계와 관련하여 검색용어 '북핵'을 이용하여 네이버트렌드 투자전략을 적용한 경우의 높은 수익률을 보여주고 있다. 이처럼 국내 주식시장에 좀 더 직접적인 영향을 주는 요인을 찾고 이러한 용어의 인터넷 검색추이와 주식시장과의 관계를 찾는 것은 앞으로 좀 더 연구해야 할 내용이라 판단된다.



[그림 6] 검색용어 '북핵'을 사용한 트렌드 투자 전략을 적용한 경우의 수익률 추이

학습기간과 검증기간 동안에 동일한 단어 집합의 수익률이 낮은 상관관계를 보이는 것과 관련해서도 마찬가지로 여러 추측이 가능하다. 먼저 동일한 검색어라 하더라도 투자와 관련하여서는 부정적 혹은 긍정적 의미를 동시에 가질 수 있다는 점을 생각해 볼 수 있다. '남북관계'라는 검색어를 예로 생각해 본다면 이것이 '남북관계 위기고조'와 같은 맥락 속에서 검색되었을 수도 있지만 '남북관계 해빙'과 같은 맥락 속에서 검색되었을 수도 있다는 것이다. 두 가지 검색 모두 '남북관계'라는 검색어를 포함하고 있지만 투자 결정과 관련해서는 서로 상반된 결과를 가져 올 것이라 예상되는 것이다. 이처럼 특정 검색어의 조회수에 따른 증감은 시간에 따라 서로 다른 맥락 속에서 사용될 가능성이 높으며, 이럴 경우 학습기간과 검증기간 동안의 수익률은 일관성을 보이기가 어려울 것이다. 추가적으로 생각해 볼 수 있는 원인으로는 축적된 데이터가 충분하지 않았을 가능성이다. 2007년 이후의 데이터만 제공되는 국내의 네이버 검색결과만으로는 충분한 학습기간과 검증기간을 가져갈 만큼의 데이터가 확보되지 않아서 두 기간의 설정과 비교자체가 완전하지 않을 수 있다는 점도 고려할 필요가 있겠다.

과연 검색어의 조합을 통해서 경제활동의 추세를

파악하여 이를 투자전략에 활용할 수 있겠는가라는 문제는 단기간의 연구만을 통해 답을 내릴 수 있는 성격의 문제는 아니다. Google의 경우 국내보다는 더 길게 2004년부터 검색결과에 대한 정보를 제공하고 있지만 이것만으로는 충분한 데이터 축적의 기간이라 할 수 없을 수도 있으며, 검색어가 나타내는 맥락이 부정적 혹은 긍정적일 수도 있다는 양면적인 특성 또한 고려해볼 필요가 있다. 최근 이러한 검색어의 증감을 실제 투자전략 수립에 활용하고 있는 기업이나 응용사례가 보고되고 있다. 만약 특정 검색어 집합이 강한 추세적 연관성을 보인다면, 이러한 사실을 활용하려는 이해당사자의 복합적인 대응관계가 오히려 이러한 연관성을 더 강화시키거나 아니면 상쇄시킬 수도 있을 것이다. 여기에 더해 각국의 상황에 따라 매우 다양한 형태로 검색어의 선택과 활용이 전개될 것이라는 점도 쉽게 예상이 가능하다.

검색어의 조합을 통한 투자전략에 대해서는 더 긴 기간에 걸친 다각적인 연구가 필요하다. 본 연구의 후속 연구로서 구글트렌드와 네이버트렌드 모두에서 높은 수익률을 보이는 검색어들과 국내의 금융시장 상황을 더 잘 표현할 수 있는 추가 검색어를 결합하여 다변량 분석을 통한 투자전략 수립에 대한 연구가 진행 중이다. 추가 검색어의 선정과 관련하여서는 국내 경제신문에 자주 언급되는 용어를 중심으로 진행되고 있으며, 국내 투자자들을 대상으로 한 조사연구도 고려중에 있다.

참 고 문 헌

[1] 료스케 지음 천재성 옮김, 빅 데이터 비즈니스, 도서출판 더 숲, 2012.

[2] 송민정, 빅 데이터가 만드는 비즈니스 미래지도, 한스 미디어, 2012.

[3] 윤형중 지음, 이제는 빅 데이터 시대, e비즈북스, 2012.

[4] 한국정보화진흥원, “신가치창출 엔진, 빅 데이터의 새로운 가능성과 대응 전략”, IT and Future Strategy, 제18호, 2011.

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

[6] Bordino, I., S. Battiston, G. Caldarelli, M. Cristelli, and Ukkonen, “Web search queries can predict stock market volumes,” *PLoS One*, Vol.7, No.7(2012), pp.1-17.

[7] Choi, H. and H. Varian, “Predicting Initial Claims for Unemployment Insurance Using Google Trends,” *Technical Report*, Google., 2009.

[8] Choi, H. and H. Varian, “Predicting the present with Google Trends,” *The Economic Record*, Vol.88(2012), pp.2-9.

[9] Cooper, C., K. Mallon, S. Leadbetter, L. Pollack, and L. Peipins, “Cancer Internet Search Activity on a Major Search Engine, United States 2001~2003,” *Journal of Medical Internet Research*, Vol.7, No.3(2005), e36.

[10] Ettredge, M., J. Gerdes, and G. Karuga, “Using Web-based search data to predict macroeconomic statistics,” *Communications of the ACM*, Vol.48, No.11(2005), pp.87-92.

[11] Ginsberg, J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, Vol.457(2009), pp.1012-1014.

[12] Goel, S., J.M. Hofman, S. Lahaie, D.M. Pennock, and D.J. Watts, “Predicting consumer behavior with Web search,” *Proceedings of the National Academy of Sciences*, Vol.7, No.41(2010), pp.17486-17490.

[13] McLaren, L. and R. Shanbhogue, “Using internet search data as economic indicator,” *Quarterly Bulletin*, Vol.Q2(2011), pp.134-140.

[14] Moat, H.S., C. Curme, A. Avakian, D.Y. Kenett, E. Stanley, and T. Preis, "Quantifying Wikipedia usage patterns before stock market moves," *Scientific Report*, Vol.3(2013), pp. 01801 : 1-5.

[15] Polgreen, P.M., Y. Chen, D.M. Pennock, and F.D. Nelson, "Using Internet Searches for Influenza Surveillance," *Healthcare Epidemiology*, Vol.47(2008), pp.1443-1448.

[16] Preis, T., D. Reith, and H.E. Stanley, "Complex dynamics of our economic life on different scales : insights from search engine query data," *Philosophical Transactions of the Royal Society*, Vol.368(2010), pp.5707-5719.

[17] Preis, T., Moat, H.S., Stanley, H.E. and Bishop, S.R., "Quantifying the Advantage of Looking Forward," *Scientific Report*, Vol.2 (2012), pp.00350 : 1-2.

[18] Preis, T., H.S. Moat, and H.E. Stanley, "Quantifying trading behavior in financial markets using Google Trends," *Scientific Report*, Vol.3(2013), pp.01684 : 1-5.