지능정보연구 제18권 제2호 2012년 6월(pp.143~156)

뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형

김유신
국민대학교 BIT전문대학원
(trust@kookmin.ac.kr)

김남규
국민대학교 경영정보학부
(ngkim@kookmin.ac.kr)

정승렬
국민대학교 BIT전문대학원
(srjeong@kookmin.ac.kr)

누구나 뉴스와 주가 사이에는 밀접한 관계를 있을 것이라 생각한다. 그래서 뉴스를 통해 투자기회를 찾고, 투자이익을 얻을 수 있을 것으로 기대한다. 그렇지만 너무나 많은 뉴스들이 실시간으로 생성·전파되며, 정작 어떤 뉴스가 중요한지, 뉴스가 주가에 미치는 영향은 얼마나 되는지를 알아내기는 쉽지 않다. 본 연구는 이러한 뉴스들을 수집·분석하여 주가와 어떠한 관련이 있는지 분석하였다. 뉴스는 그 속성상 특정한 양식을 갖지 않는 비정형 텍스트로 구성되어있다. 이러한 뉴스 컨텐츠를 분석하기 위해 오피니언 마이닝이라는 빅데이터 감성분석 기법을 적용하였고, 이를 통해 주가지수의 등락을 예측하는 지능형 투자의사결정 모형을 제시하였다. 그리고, 모형의 유효성을 검증하기 위하여 마이닝 결과와 주가지수 등락 간의 관계를 통계 분석하였다. 그 결과 뉴스 컨텐츠의 감성분석 결과값과 주가지수 등락과는 유의한 관계를 가지고 있었으며, 좀 더 세부적으로는 주식시장 개장 전 뉴스들과 주가지수의 등락과의 관계 또한 통계적으로 유의하여, 뉴스의 감성분석 결과를 이용해 주가지수의 변동성 예측이 가능할 것으로 판단되었다. 이렇게 도출된 투자의사결정 모형은 여러 유형의 뉴스 중에서 시황·전망·해외 뉴스가 주가지수 변동을 가장 잘 예측하는 것으로 나타났고 로지스틱 회귀분석결과 분류정확도는 주가하락 시 70.0%, 주가상승 시 78.8%이며 전체평균은 74.6%로 나타났다.

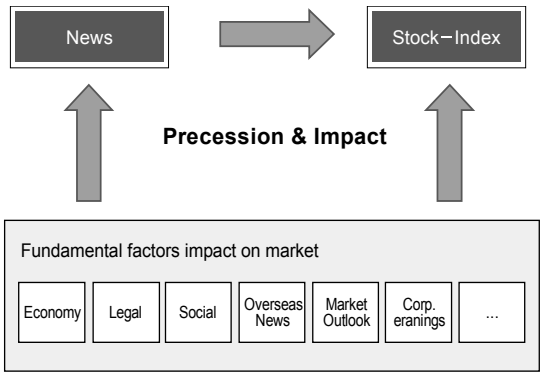
논문접수일 : 2012년 05월 16일 논문수정일 : 2012년 06월 14일 게재확정일 : 2012년 06월 18일
투고유형 : 학술대회우수논문 교신저자 : 정승렬

1. 서론

주식시장은 특유의 복잡한 가격결정 메커니즘으로 인해 주가의 변동을 시장 펀더멘탈의 변화로 설명할 수 없는 경우가 자주 발생한다. 펀더멘탈의 뚜렷한 변화가 발생하지 않았음에도 불구하고 가격이 크게 변동하는 것을 발견할 수 있는데, 이때 새로운 뉴스의 출현이 가격변동의 중요한 원인으로 종종 작용하곤 한다. 뉴스는 현실 세계에 일어나는 각종 현상에 대한 설명과 미래의 정치, 경제, 사회, 기업 등과 관련하여 앞으로 어떤 변화가 발생되고 진행되어 갈지 그에 대한 정보들을 포함하고 있기 때문이다. 그러므로 뉴스와 주가는 밀접한 관계를 가지고 있으며, 뉴스를 통해 시장 참가자들은 주식시장의 변동성을 일부나마 예측할 수 있게 된다(송치영, 2002).

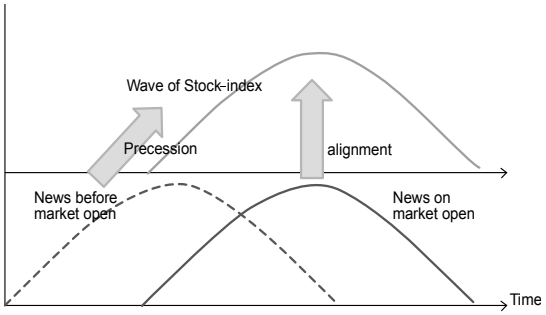
하지만 <Figure 1>과 같이 주가에 영향을 미치는 펀더멘털 요인들은 너무나도 다양하고 복잡하

며 이러한 요인들이 뉴스와 주가에 영향을 미치고 뉴스는 다시 주가에 영향을 미치는 식의 순환이 발생하기도 한다.



<Figure 1> Fundamental Factors on Stock Market

결국 뉴스는 <Figure 2>에서 보여지듯 주가에 영향을 미치는 영향 요인이 되기도 하고 주가의 흐름을 미리 보여주는 선행지표가 되기도 한다.



<Figure 2> Influence of News to Stock-price

그러나 하루에도 수없이 많은 뉴스들이 나타나고 사라지고 있어, 뉴스를 하나하나 분석하여 주가에 미치는 영향을 파악하기란 거의 불가능한 일이다. 더군다나 거시적 관점의 정책, 전망뉴스부터 매일 매일의 시황 · 실적 · 기업뉴스 등 다양한 유형의 뉴스들이 실시간으로 양산되며, 그 내용이 시장에 긍정적인지 부정적인지 명확히 파악하기가 쉽지 않다. 또한 뉴스라는 속성상 다소 중립적인 뉘앙스로 주식시장의 긍정/부정 양쪽 의견을 모두 제시하는 경우가 많기 때문에 실상 그 뉴스의 저의를 파악하는 것 또한 간단치 않으며, 뉴스를 분석하는 사람마다의 주관에 따라 달라질 위험성이 존재한다.

때문에 기존의 연구들 역시 쉽게 판별이 가능한 특정 사건과 뉴스들을 위주로 그에 반응하는 주가, 또는 주가가 크게 변동되었을 때 이를 야기한 뉴스가 존재했는지를 역으로 분석하곤 하였다. 그러나 뉴스는 대부분 일정한 양식이나 속성이 없는 텍스트들로 구성되어 있으며, 하루에도 수없이 많은 뉴스들이 양산된다. 최근 이러한 대량의 데이터를 분석하여 의미 있는 정보로 변환하고자 하는 빅데이터 분석이 주목 받고 있으나 아직은 초기 단계라 할 수 있다. 본 연구에서는 빅데이터 분석 기법 중 오피니언 마이닝을 이용하여 비정형 뉴스 컨텐츠의 감성을 분석하고 수치화하여 뉴스가 주식 시장에 미칠 영향을 분석하였으며, 이를 통해 투자정보를 생성하는 지능형 투자의사결정 모형을 제시하고자 하였다. 또한 이러한 모형의 핵심 근간인 뉴스와 주가간의 관계를 실증 분석함으로써 뉴스 컨텐츠가 오피니언 마이닝을 통해 투자정보로 제시될 수 있는지 확인해 보고자 한다.

본 연구는 다음과 같이 구성되어 있다. 제2장에서는 뉴스와 주가에 대한 선행연구와 빅데이터 분석 및 오피니언 마이닝에 대해서 소개하고, 제3장에서는 지능형 투자의사결정 모형의 구조 및 처리 프로세스와 방법을 설명하고, 뉴스와 주가간의 관계 분석을 위한 실증분석 모형을 제시한다. 제4장에서는 실증분석 결과를 기술하고, 제5장에서 결론 및 향후 연구에 대하여 논의한다.

2. 관련 연구

2.1 주가지수 예측

예로부터 주가 예측을 위한 많은 연구들이 있었다. 전통적으로 이러한 연구들은 주로 시계열분석 또는 펀더멘털 관점에서의 통계적 분석을 시도하였다. 시계열분석은 주가자체가 모든 정보를 반영한다는 믿음에 기반하며, 펀더멘털분석은 주가가 이자율 등의 거시경제 변수에 영향을 받는다는 것이다. 그러나 전통적 방식으로는 만족할만한 예측 결과를 얻지 못하였고, 이후 새롭게 등장한 분석기법으로 인공지능(AI : Artificial Intelligence)을 이용한 주가예측이 시도되었다. 박종엽, 한인구(1995)는 인공신경망을 이용한 한국종합주가지수(KOSPI)의 방향성 예측을 통해 다중회귀분석(MLR)보다 높은 정확도를 확인하였다.

이후에도 인공지능을 이용한 주가지수 예측 노력은 계속되었고, 트레이딩 알고리즘을 찾아내려는 연구가 활발히 진행되었다. 김선웅, 안현철(2010)은 기존 연구의 대다수가 가격기반의 기술적 지표를 주 분석대상으로 하며 실제 트레이딩의 수익률보다는 주가 등락의 예측 정확도에만 한정되어 있음에 착안하여, 기술적 지표 외에 주가에 영향을 미치는 다양한 비가격 변수를 추가하고 주가예측의 성과를 누적 투자수익률 관점에서 분석하였다.

한편 김진화 등(2011)은 지식 누적을 이용한 실시간 주식시장 예측 연구를 통해 주가가 연속발생 데이터라는 점에서 마이닝 대상 자료의 변화를 다루는 점진적 데이터마이닝 분석을 시도하였다. 이는 계속 증가되는 데이터를 반영할 수 있도록 예측모델을 갱신함으로써 예측 정확도를 높이고자 하는 것을 포함하고 있다.

이러한 기술적 지표 위주의 주가 분석 및 예측 연구와는 달리 뉴스와 주가간의 관계를 분석하고 주가를 예측하고자 하는 연구들도 다수 진행되었다. Mark(1994)는 뉴스와 주가변동성 간의 관계가 존재함을 확인하였고, Veronsei(1999)는 호황기 부정적인 뉴스에 주식시장이 반응하는 정도에 대한 연구를 실시하였다. Lee Gillam(2002)은 뉴스의 Good/Bad 빈도를 측정하여 FTSE 100 지수와 비교하였고, Tak-chung Fu(2008)는 홍콩주식시장을 대상으로 주가흐름과 뉴스 사이의 핵심연계지점을 도출하고자 하였다.

국내에서는 송치영(2002)이 주가의 일별 변동폭이 큰 200개의 표본을 추출하여 뉴스와 비교하여, 뉴스의 발생이 주가에 중요한 영향을 미쳤으며 뉴스 유형별로 유리한 뉴스와 불리한 뉴스에 주가가 다르게 반응하는 비대칭적 반응을 보이고 있음을 실증 분석하였다. 이근영(2006)은 전쟁위험의 증대가 주식시장의 주가를 하락시킨다고 하였고, 안희준(2010)은 남북관계에서 일어나는 특정 사건들과 주식시장과의 상관관계를 분석한 결과 뉴스가 주가에 유의미한 영향을 미친 것으로 분석하였으며, 신종화, 이의철(2010)은 2000년 들어 등장한 인터넷 뉴스 매체를 통해 실시간으로 전파되는 뉴스들이 시장에 매우 큰 영향을 미치고 있음을 여러 가지 사례를 통해 보여주고 있다. 안성원, 조성배(2010)는 뉴스 텍스트에서 추출한 특성을 주가의 변동폭과 대비하여 Naïve Bayesian 분류기로 학습 후 개별 종목 주가의 상승/하락을 예측하는 모형을 제시하였다.

이와 같은 선행연구들은 뉴스가 주가에 영향을 미치고 있음을 분명히 제시하고 있으나, 주로 특정 뉴스나 특정 종목의 뉴스를 대상으로 하고 있어 하루에도 수많은 뉴스들이 양산되고 실시간으로 전파되고 있는 현실 세계의 즉시적 반영과 분석은 미흡한 실정이다.

2.2 빅데이터(Big Data) 분석

현대 사회는 뉴스뿐만 아니라 다양한 분야에서 무수히 많은 데이터가 실시간으로 발생하고 있다. 특히 스마트 단말 및 소셜미디어 등으로 대표되는 다양한 정보채널의 등장과 이로 인한 정보의 생산, 유통, 보유량의 증가는 계속적으로 데이터의 기하급수적인 증가를 이끌고 있다(조성우, 2011).

하지만 급증하고 있는 데이터는 기존의 관리 및 분석 체계로는 감당할 수 없어 빅데이터 분석이 필요하다(채승병, 2011). 그러나 빅데이터라는 용어 자체가 다분히 추상적이고 광범위하기 때문에 이를 명확히 정의하기도 쉽지 않고 이에 대한 연구와 활용도 미미하다.

빅데이터의 전통적 개념은 구글과 같은 대기업이나 NASA의 연구과학 프로젝트에서 분석하는 대용량의 데이터를 일컫는 것이었다(Nerv Adrian, 2011). Manovich(2011)는 시간 흐름상 데이터를 수집하고 처리하던 소프트웨어 도구의 능력을 넘어서는 데이터들의 모임을 빅데이터로 정의하고, 이런 데이터의 크기는 앞으로 꾸준하게 그 범위가 변화할 것이라고 하였다. McKinsey에서는 빅데이터를 수집 · 저장 · 소통 · 집단화 · 분석이 가능한 거대한 데이터 풀(Pool)로서 정의하고, 이제는 글로벌 경제의 모든 영역과 기능의 일부가 되었다고 하였다(McKinsey and Company, 2011).

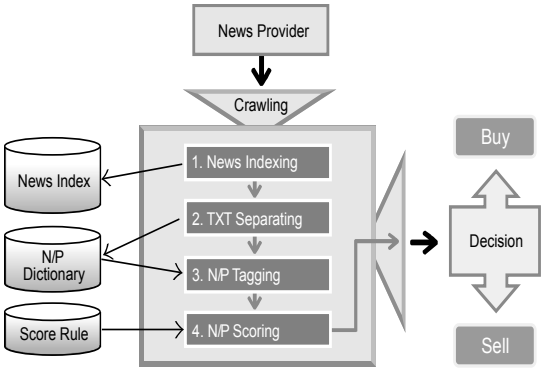
이러한 빅데이터를 분석하는 다양한 기법들이 있지만, 최근 소셜미디어 등 비정형 데이터의 증가로 인해 텍스트 마이닝, 오피니언 마이닝, 소셜네트워크 분석, 군집 분석 등이 주목을 받고 있다(조성우, 2011).

3. 지능형 투자의사결정모형

본 연구에서 제시하는 지능형 투자의사결정 모형은 실시간으로 쏟아져 나오는 대량의 뉴스 컨텐츠를 오피니언 마이닝으로 분류 · 분석하여 주식시장에의 영향도를 평가하고 투자의사결정 정보를 제공하는 빅데이터 감성분석 시스템 모형이다.

3.1 지능형 투자의사결정 모형 구조

빅데이터 감성분석을 통한 지능형 투자의사결정 모형의 구조는 <Figure 3>과 같다



<Figure 3> Intelligent Stock-index Investment Model

투자의사결정 모형의 첫째 단계는 실시간으로 뉴스를 수집하는 뉴스 프로바이더로부터 뉴스를 공급받고 뉴스의 정보를 인덱싱하는 것이다. 뉴스의 내용뿐만 아니라 매체, 시간, 뉴스 유형 등 다양한 정보들을 수집 · 분류하여 투자의사결정을 유추할 수 있는 변수로서 재가공한다. 다음 단계는 뉴스 컨텐츠의 텍스트를 형태소로 분리하여 극성을 판단할 수 있는 단어를 도출하고 이를 긍정/부정 사전과 비교하여 단어 별 긍정/부정 극성을 태깅한다. 세 번째는 인덱싱 된 분류정보와 스코어링 룰을 이용하여 뉴스의 긍정/부정 극성을 판별하고 이를 다시 일별 스코어링 기준에 의해 최종 투자의사결정 정보를 도출한다. 뉴스의 긍정/부정 극성을 판별하고 계산하는 식은 다음과 같다.

1) 먼저, 1개의 뉴스에 대한 긍정/부정 비율 NewsPNr을 계산한다.

$$\text{NewsPNr} = \frac{\sum_{i=1}^{n} wordPN(i)}{n} \times 100\%$$

① 하나의 뉴스에는 여러 단어가 존재하며, 각 단어는 긍정/부정의 극성을 가진다.
② 각 단어를 긍정/부정 감성 사전과 비교하여 극성을 태깅한다.
WordPN(i) = 단어의 긍정/부정값(1 or 0)
③ 각 단어들의 긍정/부정 값을 뉴스의 긍정/부정 비율로 합산한다.

2) 뉴스에 대한 긍정/부정 판별 NewsPN(P, N, NEU) 조건은 다음과 같다.
if(NewsPNr > 52), NewsPN = Positive;
else if(52 > = NewsPNr > = 48), NewsPN = Neu;
else if(NewsPNr < 48), NewsPN = Negative

3) 하루 동안 발생한 모든 뉴스의 긍정/부정을 도출하는 식은 다음과 같다.
DayPNr = (NewsPNr1+NewsPNr2+ … +NewsPNr*n*)/*n* $= \frac{\sum_{j=1}^{n} NewsPNr(j)}{n}$

① 뉴스의 인덱싱 정보와 스코어링 룰에 따라 다양한 변수들로 재가공된다.

$$\text{Day.X.PNr} = \frac{\sum_{j=1}^{n} News.X.PNr(j)}{n}$$

② 하루 동안 발생한 전망(forecast) 뉴스들의 긍정/부정 스코어는 다음과 같다.

$$\text{Day.fcPNr} = \frac{\sum_{j=1}^{n} News.fcPNr(j)}{n}$$

4) 투자의사결정정보는 투자의사결정 회귀함수 $\hat{Y}$를 통해 Buy/Sell 의견을 제시한다.

$$\hat{Y} = b0+b1X1+b2X2+ \cdots +bnXn$$

3.2 오피니언 마이닝 결과와 주가 비교

본 연구에서 제시한 지능형 투자의사결정 모형은 뉴스 컨텐츠의 분석결과를 투자정보로 변환하여 투자의사결정을 지원하는 것이다. 이는 곧 뉴스와 주가간의 관계가 명확해야 함을 의미한다. 그러므로 시스템의 유효성을 검증하기 위해 뉴스 컨텐츠의 오피니언 결과와 주가와의 관계를 통계기법을 이용하여 실증 분석하고자 한다. 이를 위해 다음과 같은 가설을 제시하고 검증한다.

- 뉴스의 긍정/부정 의견은 주가에 영향을 미칠 것이다.
- 뉴스의 긍정/부정 의견 보다 긍정/부정 비율이 주가변동을 더 잘 설명할 것이다.
- 뉴스의 유형에 따라 주가에 미치는 영향이 다를 것이다.

첫째 가설을 살펴보기 위해 뉴스의 긍정/부정 오피니언을 평가하고 이를 주식시장 개장 전후로 구분하여 뉴스의 긍정/부정 의견과 주가 등락 사이에 유의한 관계가 있는지를 분석한다. 또한 다양한 뉴스의 유형이 각기 주가지수 등락과 관계가 있는지를 분석한다.

둘째 가설은 본 모형이 오피니언 마이닝으로 뉴스의 긍정/부정 의견을 도출하지만, 긍정 또는 부정이라는 단순 평가보다는 뉴스가 가진 긍정/부정 비율을 계산하여 이를 활용한다면 좀 더 세밀한 비교가 가능할 것으로 예측되어 이를 검증하고자 한다. 본 가설이 맞는다면 긍정부정 비율을 이용할 때 주가지수와의 관계를 더 잘 예측할 수 있을 것이다.

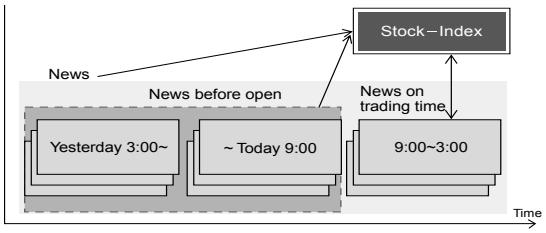
마지막 가설은 여러 가지 유형의 뉴스가 존재하지만 모든 뉴스가 주가지수의 변동성을 설명할 수 있는 것은 아닐 것이며, 특별히 예측력이 높은 뉴스 유형이 존재할 것으로 생각되어 이를 분류하고자 하는 것이다. 이를 통해 투자의사결정에 필요한 요인을 도출할 수 있을 것으로 예상된다.

3.3 실증분석 설계

본 연구를 위해 한국거래소에서 2011년 7월부터 9월까지 3개월간 주식시장이 개장된 63일의 KOSPI 지수와 지수등락폭을 수집하였고, 뉴스 데이터는 포털 사이트인 네이버의 증권정보 > 뉴스 > 주요뉴스에 게재된 기사 중 실시간 경제뉴스 매체 M사의 기사 766건을 웹 페이지에서 파싱하여 수집하였다.

뉴스가 발생한 시점이 주식시장 개장중일때와 그렇지 않을 때 미치는 영향이 다를 수 있음을 감안하여 <Figure 4>와 같이 뉴스의 시간별 구분을 하였고, 주식시장 개장인 오전 9시를 기준으로 전일 장마감 후부터 당일 오전 9시 까지를 “장전뉴스”, 9시부터 오후 3시까지는 “장중뉴스”로 구분하고, “뉴스”는 “장전뉴스”와 “장중뉴스”를 포함하였다.

뉴스 텍스트 마이닝을 위한 형태소 분석과 극성 분류는 워즈워드 seHANA SW를 사용하였고, 수집된 데이터의 통계 분석은 IBM-PASW 18을 이용하였다.



<Figure 4> Analysis Model of News and Stock-index

4. 실증 분석 결과

2011년 7월부터 9월까지 3개월 동안 주식시작은 휴일을 제외한 63일이 개장되었고, 그 중 상승 33일 하락 30일의 변동이 있었다. 동 기간 내 수집된 뉴스 컨텐츠는 총 766건이며, 주식시장 운영 시간 기준으로 구분한 개장 전 뉴스는 197건, 장중 뉴스는 385건, 장 마감 후 뉴스는 184건으로 각 유형별 개수와 시간구분 별 합계와 비율은 <Table 1>과 같다. 이중 개장 전 뉴스와 어제 장마감 후 뉴스는 장전 뉴스로 합산되었다.

뉴스의 긍정/부정 개수와 비율은 아래 <Table 2>와 같다. 조사기간 내 주가가 상승한 날은 33일로 하락한 30일보다 많았음에도 불구하고 뉴스의 60%가 부정의견을 보였고, 단지 20%만이 긍정의견으로 판명되었다. 이는 극명하게 시장이 좋지 않은 한 뉴스는 대체로 보수적 관점의 조심스러운 태도를 보이고 있음을 알 수 있다. 그러므로 뉴스를 긍정이냐 부정이냐의 단선적 의견으로 판단할 경우 주가의 흐름을 설명하는데 다소 비약이 존재할 수 있음을 유추해 볼 수 있다.

<Table 1> Frequency of News by Nature(Positive vs. Negative) of News by Time

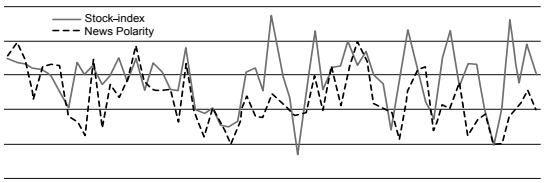
구 분		Economy	Condition	Industry	Outlook	Corporate	Oversea	Sum
News		**77**	**264**	**70**	**186**	**71**	**98**	**766**
	Before	24	2	14	54	14	89	197(26%)
	Open	37	210	29	67	36	6	385(50%)
	After	16	52	27	65	21	3	184(24%)

<Table 2> Frequency of News by Nature(Positive vs. Negative) of News by Polarity

구 분		Economy	Condition	Industry	Outlook	Corporate	Oversea	Sum
News		**77**	**264**	**70**	**186**	**71**	**98**	**766**
	Neg	47	152	40	110	36	71	456(60%)
	Neu	8	30	5	21	3	11	78(10%)
	Pos	22	82	25	55	32	16	232(20%)

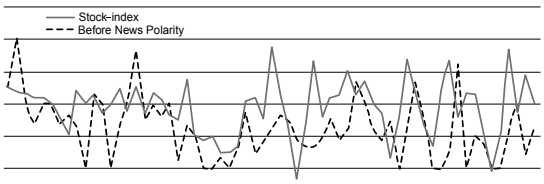
4.1 뉴스와 주가의 변동 곡선

뉴스의 긍정/부정 의견과 주가지수 등락이 서로 유사한 패턴을 보이는지 살펴보기 위해 조사기간 동안의 주가지수 등락의 흐름과 일자별 뉴스의 긍정/부정 마이닝 결과를 수치로 환산하여 그래프로 비교해 보았다. 아래 <Figure 5>에서 보이는 바와 같이 두 그래프가 매우 유사한 흐름을 보이고 있음을 알 수 있다. 이는 대체로 긍정의견이 많을 때 주가가 오르고, 부정 의견이 많을 때 주가가 하락하고 있음을 보여주는 것이라 하겠다.



<Figure 5> Flow of Stock-index and News

<Figure 5>는 개장된 상태의 뉴스를 포함하고 있어 주가 흐름과 유사한 형태를 보이는 것이 당연하다고 할 수 있다. 하지만 개장 전 뉴스와 주가의 흐름을 비교한 <Figure 6>에서도 뉴스의 긍정/부정 마이닝 결과가 주가 흐름과 유사함을 알 수 있다.



<Figure 6> Flow of Stock-index and Before News

이러한 주가지수 등락과 뉴스의 긍정/부정 패턴이 유사함은 곧 뉴스와 주가가 매우 밀접한 관계를 가지고 있으며, 특히 장전뉴스가 주가에 일정수준 선행하고 있거나 영향을 미치고 있다고 볼 수 있을 것이다.

4.2 뉴스와 주가 등락

4.2.1 시간 기준별 뉴스와 주가등락 비교

앞서 주가와 뉴스의 흐름을 그래프로 살펴보았던 것을 주식시장 전후 시간별로 구분하여 이때의 긍정/부정 마이닝 결과가 주가의 등락(Up/Down) 여부와 유의한 관계를 가지는지 독립표본 T검정을 실시한 결과는 <Table 3>과 같다.

<Table 3> KOSPI Up/Down by Time

		KOSPI Up Pos-Avg.	KOSPIDown Pos-Avg	p-value
News PN	Overall	-18.45	-50.21	.001*
	Before	-26.32	-50.65	.036**
	Open	-9.09	-50.54	.002*
News PNr	Overall	.4833	.3757	.002*
	Before	.4671	.3820	.005*
	Open	.4998	.3523	.000*

주) *, ** : 각각 99%, 95% 수준(2-tailed)에서 통계적으로 유의함.
- 뉴스PN : 뉴스의 긍정/부정 극성을 100과 -100으로 단순화한 값.
- 뉴스PNr : 뉴스의 긍정/부정 극성을 비율(Ratio)로 표현한 값.

뉴스는 시간 기준별 구분과 무관하게 주가 등락과 유의한 것으로 분석되었으며, 뉴스의 극성 평균에 따르면 주가가 오르던 내리던 관계없이 뉴스의 기조는 대체로 부정적이며 주가 상승 시 덜 부정적인 것을 알 수 있다.

4.2.2 뉴스 유형과 주가등락 비교

다음으로 유형별 뉴스가 주가의 등락과 유의한 관계를 가지는지 살펴보았다. 뉴스 유형에는 환율의 변화, 경기 흐름, 정부 정책 등의 경제뉴스를 비롯하여 지금 당장의 주식시장 흐름을 보여주는 시황 뉴스, 주가의 상승/하락을 예측하는 전망 뉴스, 경제 여건에 따라 영향을 받거나 특별하게 언급될 만한 업종 뉴스,

개별 기업의 성과 등을 알리는 종목 뉴스, 미국 · 유럽 등의 해외시장에 대한 소식을 전하는 해외뉴스 등이 있다. 이러한 뉴스 유형에 따라 주식시장에 미치는 영향이나 그 관계의 유의성이 다를 것으로 예상되어 주가의 등락과 비교하여 T검정을 실시하였다.

<Table 4> T-test of KOSPIUp/Down by News Type

		KOSPI Up Pos-Avg.	KOSPIDown Pos-Avg.	p-value
Economy	Before	.4432	.4760	.625
	Open	.4390	.4834	.495
Condition	Before	.5758	.3281	.000*
	Open	.5438	.3246	.000*
Industry	Before	.5053	.3575	.066
	Open	.4987	.3754	.111
Outlook	Before	.4946	.3683	.000*
	Open	.5081	.3720	.000*
Corporate	Before	.5015	.6411	.192
	Open	.5452	.6149	.469
Oversea	Before	.4100	.2700	.006*
	Open	.4097	.2816	.011**

주가 등락과 유형별 뉴스의 마이닝 결과값을 검증 한 결과 <Table 4>와 같이 시황 · 전망 · 해외뉴스 만이 주가 등락과 유의한 관계를 가지는 것으로 나타났다. 또한 뉴스들은 대체로 부정적인 의견을 나타내고 있지만, 주가가 오를 땐 다소 중립적인 의견을 보이고 주가가 하락할 때는 매우 부정적인 의견을 나타내고 있음을 알 수 있다. 한편, 개별 기업 종목 뉴스와 같은 경우는 하락시장에서의 오피니언 긍정 값이 높게 나타났는데, 이는 주가하락을 방어하고자 하는 긍정적 내용의 홍보기획 기사들이 다수 출현했기 때문으로 유추해 볼 수 있을 것이다.

4.3 뉴스의 극성 판별과 비율 값의 설명력 비교

오피니언 마이닝에서는 분석결과를 긍정 또는 부정 의견으로 단순 판별하는 경향이 있으나, 본 연구에서는 뉴스 컨텐츠를 긍정/부정 의견으로 단순화하기 보다는 긍정/부정을 비율로 산출하여 분석할 때 더 높은 설명력을 가질 것으로 예상하였다. 이에 대한 비교 결과 <Table 5>와 같이 긍정/부정 비율로 분석하였을 때 주가변동에 미치는 영향에 대한 설명력이 더 높은 것으로 나타났다.

<Table 5> R^2 of PNr more than PN

		R	R^2	p-value
PN	News PN	.481	.231	.000*
	Before	.274	.075	.030**
	Open	.472	.223	.000*
PNr	News PNr	.604	.364	.000*
	Before	.366	.134	.003*
	Open	.544	.296	.000*

4.4 유형별 뉴스와 주가지수 비교

주가지수 변동을 예측하기 위해서 어떤 유형의 뉴스를 변수로 사용해야 할 것인지, 주가지수와의 관계와 설명력은 어느 정도인지 알아보기 위해 유형별 뉴스와 주가지수를 대상으로 다중회귀분석을 실시하였다.

분석결과 모든 유형의 뉴스를 대상으로 주가를 설명하고자 했을 때의 회귀식은 유효하지 않은 것으로 나타났다. 결과적으로 앞서 분석한 뉴스 유형과 주가등락 T검정에서 유의하였던 시황· 전망·해외뉴스만을 변수로 하였을 때의 회귀식이 신뢰수준 99%에서 통계적으로 유의하였으며, <Table 6>에서 보여지듯 총 변동에 대한 설명력도 뉴스에서 52.7%, 장전뉴스만을 대상으로 분석하였을 때에도 51.8%로 비교적 높게 나타났다. 또한 시황 · 전망 · 해외뉴스 변수 사이의 VIF가 3 이하로 다중공선성에도 문제가 없는 것으로 분석되었다. 결국, 뉴스 유형 중 시황 · 전망 · 해외뉴스는 투자의사결정 함수에 적용할 수 있는 주요 변수임을 의미하는 것이라 하겠다.

<Table 6> Results of Multiple Regression

		R	R^2	F	p-value	-
		-	B	t	p-value	VIF
Before PNr		.719	.518	16.456	.000*	-
	Condition	-	76.787	3.248	.002*	1.364
	Outlook	-	92.705	2.351	.023**	1.306
	Oversea	-	59.106	2.170	.035**	1.288
Open PNr		.726	.527	17.109	.000*	-
	Condition	-	70.114	2.598	.013**	1.453
	Outlook	-	122.819	2.854	.006*	1.434
	Oversea	-	60.073	2.291	.027**	1.222

4.5 주가지수 등락 예측 회귀식

마지막으로 주가지수등락을 예측하기 위한 투자 의사결정함수를 도출하고자 일자 별 장전 시황・전망・해외뉴스의 긍정부정 비율을 이용하여 로지스틱 회귀분석을 실시하였다. 회귀식의 -2 Log 우도는 55.330이였으며, 모형의 적합도 검정 Hosmer와 Lemeshow 검정의 유의확률 값이 .204로 α = 0.05보다 크므로 모형이 부적합하다는 귀무가설을 기각하였고, 독립변수 모두 통계적으로 유의하여(유의확률 = .000) 장전 시황・전망・해외뉴스가 주가에 영향을 미치는 회귀모형은 유용하다고 할 수 있다.

$$\hat{Y} = b0+b1X1+b2X2+b3X3$$

$X1$ = Day개장전. 시황뉴스PNr×100

$X2$ = Day개장전. 전망뉴스PNr×100

$X3$ = Day개장전. 경제뉴스PNr×100

×100은 비율r을 100점단위로 변환함.

예측 결과의 분류정확도는 <Table 7>과 같이 주가하락에 대해서 70.0%, 주가상승에 대해서는 78.8%로 주가하락보다는 주가상승에 대한 분류정확도가 조금 더 높았으며, 전체적으론 74.6%의 정확도를 보여주고 있다.

<Table 7> Predict Accuracy

	Predicted		
	Down	Up	Correct %
Down	21	9	70.0
Up	7	26	78.8
Overall percentage			74.6

그러나, 회귀모형의 적합도와는 별개로 각 독립변수의 회귀계수에 대한 유의확률을 통계적 유의수준 α = 0.05과 비교해보면 <Table 8>에 보여지는 바와 같이 장전시황(.003)과 장전전망(.023) 뉴스만이 통계적으로 유의한 것으로 나타났다.

<Table 8> Logistic Regression Analysis

Before	B	p-value	Exp(B)
Condition	.073	.003*	1.076
Outlook	.075	.023**	1.078
Oversea	.001	.977	1.001
Constant	-6.492	.000	.002

결국, 주가지수의 등락을 예측하기 위한 투자의사결정함수 $\hat{Y}$는 다음과 같다.

$$\hat{Y} = -6.492 + .073X1(\text{장전. 시황뉴스}) + .075X2(\text{장전. 전망뉴스}) + .001X3(\text{장전. 해외뉴스})$$

5. 결론

5.1 연구 결과 요약 및 논의

본 연구는 뉴스와 주가가 밀접한 관계를 가지고 있을 것 이라는 가정에서 출발하여 뉴스와 주가의 관계를 살펴보고, 뉴스를 분석하여 투자정보를 추출하고자 하였다. 뉴스는 그 속성상 비정형 텍스트로 구성되어 있으며 하루에도 수없이 많은 뉴스가

실시간으로 양산되고 있다. 대량의 뉴스를 시스템적으로 수집 · 분류 · 분석하여 투자정보를 생성하기 위해 빅데이터 분석기법 중 오피니언 마이닝분석을 이용하였고, 이를 기반으로 지능형 투자의사결정 모형을 제시하였다. 그리고, 모형의 유효성을 증명하기 위해 뉴스 오피니언 마이닝 결과와 주가지수 간의 관계를 통계기법을 이용하여 실증 분석하였다.

그 결과 뉴스 컨텐츠의 긍정/부정 오피니언과 주가 등락과는 유의한 관계를 가지며, 긍정/부정의 단순화된 의견으로 판별되는 뉴스 오피니언을 긍정/부정 비율로 도출하여 적용할 때 주가지수의 흐름을 더 잘 설명하는 것으로 나타났다. 또한 뉴스가 주가변동에 영향을 미치거나 적어도 선행하고 있는가를 확인하기 위해 주식시장 개장 전 발생한 뉴스들만으로 주가흐름과 비교한 결과 역시 통계적으로 유의한 것으로 확인되었다.

또한 뉴스에는 사회 · 경제, 해외소식, 기업실적, 업종현황, 시장전망, 시장현황 등 다양한 형태와 정보를 내포하고 있는 만큼 뉴스 유형에 따라 주식시장에 미치는 영향이나 그 관계의 유의성이 다를 것으로 예상되어 유형별 뉴스와 주가의 등락을 비교한 결과 시황 · 전망 · 해외 뉴스가 주가변동을 설명하는 데 가장 유용한 것으로 나타났다. 반대로 개별 기업종목에 대한 뉴스는 통계적으로 유의하지 않았지만 오피니언 마이닝 값이 주가와 반대의 흐름을 보였는데, 이는 하락 장에서 기업주가의 하락을 방어하기 위한 홍보성 뉴스들의 등장 때문으로 해석해 볼 수 있다.

마지막으로 뉴스의 긍정/부정 의견과 주가의 관계를 기반으로 한 투자의사결정 함수를 도출하고자 중회귀분석과 로지스틱 회귀분석을 실시하였고, 그 결과 주식시장 개장 전 시황 · 전망 · 해외 뉴스를 변수로 한 회귀식이 통계적으로 유의하였으며, 각 변수의 회귀계수로는 시황뉴스와 전망뉴스가 통계적으로 유의함을 알 수 있었다.

본 연구는 뉴스와 주가의 관계를 뉴스 컨텐츠의 오피니언 마이닝을 통해 최초로 해석하였고, 더 나아가 시스템적으로 오피니언 마이닝을 수행하고 투자정보를 도출하여 지원할 수 있는 지능형 투자의사결정모형을 제시하고 이를 검증하였다. 이는 뉴스를 주가지수 투자예측의 변수로서 활용할 수 있음을 보여주는 것이며, 향후 시스템으로 구현되고 검증된다면 실물투자 지원시스템으로서 활용될 수 있을 것으로 기대된다.

5.2 연구의 한계점 및 연구 방향

본 연구를 위해 수집한 3개월간의 뉴스 800여건은 KOSPI 대비 63일의 주가에 해당되어, 뉴스와 주가를 비교하기에는 데이터가 충분하지 못하였으며, 경제 · 주식시장에 특화된 도메인 감성 사전을 구축하지 못하고 범용 감성 사전을 이용함으로써 긍정/부정 극성 분류의 정확도가 다소 낮을 수 있다는 한계를 가지고 있다. 또한 개별종목의 주가가 아닌 종합주가지수를 비교의 대상으로 하고 있어, 기업에 대한 투자의사결정 정보를 제공하지 못하는 한계가 있다.

그럼으로 향후 연구에서는 주식시장에 특화된 도메인 감성사전을 구축하고, 충분한 양의 데이터를 수집하여 분석한다면 지능형 투자의사결정모형의 완성도를 높일 수 있을 것이다. 또한 모형을 실제 지능형 투자의사결정 시스템으로 구현하고 도출된 투자정보를 선물 · 옵션 투자와 같은 실물투자에 시뮬레이션 해봄으로써 투자성과를 확인하고 조정하여 모형의 예측력과 정확도를 높일 수 있을 것으로 기대된다.

참고문헌

김선웅, 안현철, “Support Vector Machines와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발”, *지능정보연구*, 16권 1호, 2010.

김진화, 홍광헌, 민진영, “지식 누적을 이용한 실시간 주식시장 예측”, *지능정보연구*, 17권 4호(2011).

박종엽, 한인구, “인공신경망을 이용한 한국종합주가지수의 방향성 예측”, *한국전문가시스템학회지*, 2호(1995).

송종석, 이수원, “상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축”, 정보과학회 논문지 : 소프트웨어 및 응용, 38권 3호(2011).

송치영, “뉴스가 금융시장에 미치는 영향에 관한 연구”, *국제경제연구*, 8권 3호(2005).

신종화, 이의철, “한국형 경제 뉴스 속보가 금융 시장에 미친 영향”, 한국경영학회 통합학술대회, 2010.

안성원, 조성배, “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측”, 2010 한국컴퓨터 종합학술대회 논문집, 37권 1(C)호(2010).

안희중, 전승표, “남북관계 관련 뉴스가 주식시장에 미치는 영향”, 한국금융연구원 한국경제의 분석, 16권 2호(2010).

이근영, “북한 핵 관련 뉴스가 국내주식 및 외환시장에 미치는 영향”, *동북아경제연구*, 18권 1호(2006).

이완수, 김찬석, “TV 기업뉴스와 주식가치의 상관관계 대한 시계열 연구”, 방송학회학술대회 논문집, 2009.

윤홍준, 김한준, 장재영, “오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법”, 정보과학회 논문지 : 컴퓨팅의 실제 및 레터, 16권 2호(2010).

조성우, “Big Data 시대의 기술”, KT종합기술원, 2011.

채승병, “정보홍수 속에서 금맥 찾기 : 빅데이터 분석과 활용”, SERI 경영노트, 91호(2011).

James, Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers, “Big Data : The Next Frontier for Innovation, Competition, and Productivity”, *McKinsey and Company, The McKinsey Global Institute*, 2011.

Lee, Gillam, Khurshid Ahmad, Saif Ahmad, Matthew Casey, David Cheng, Tugba Taskaya, Paulo C F de Oliveira and Pensiri Manomaisupat. “Economic News and Stock Market Correlation : A Study of the UK Market”, *In Conferenceon Terminology and Knowledge Engineering*, 2002.

Manovich, L., “Trending : The Promises and the Challenges of Big Social Data”, *Debatesinthe Digital Humanities*, 2011.

Mark, L. Mitchell and J. Harold Mulherin, “The Impact of Public Information on the Stock Market”, *The Journal of Fiance*, Vol.XL IX, No.3(1994).

Merv Adrian, “이제는 빅데이터이다”, *Teradata Magazine*, 2011.

Pietro Veronesi, “Stock Market Overreaction to Bad News in Good Times : A Rational Expectations Equilibrium Model”, *The Review of Financial Studies*, Vol.12, No.5(1999).

Tak-shung Fu, Ka-ki Lee, Donahue Sze, Fu-lai Chung, and Chak-man Ng, “Discovering the Correlation between Stock Time Series and Financial News”, *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligenceand Intelligent Agent Technology*, 2008.

Abstract

Stock-Index Invest Model Using News Big Data Opinion Mining

Yoosin Kim* · Namgyu Kim** · Seung Ryul Jeong***

People easily believe that news and stock index are closely related. They think that securing news before anyone else can help them forecast the stock prices and enjoy great profit, or perhaps capture the investment opportunity. However, it is no easy feat to determine to what extent the two are related, come up with the investment decision based on news, or find out such investment information is valid. If the significance of news and its impact on the stock market are analyzed, it will be possible to extract the information that can assist the investment decisions. The reality however is that the world is inundated with a massive wave of news in real time. And news is not patterned text.

This study suggests the stock-index invest model based on "News Big Data" opinion mining that systematically collects, categorizes and analyzes the news and creates investment information. To verify the validity of the model, the relationship between the result of news opinion mining and stock-index was empirically analyzed by using statistics.

Steps in the mining that converts news into information for investment decision making, are as follows. First, it is indexing information of news after getting a supply of news from news provider that collects news on real-time basis. Not only contents of news but also various information such as media, time, and news type and so on are collected and classified, and then are reworked as variable from which investment decision making can be inferred. Next step is to derive word that can judge polarity by separating text of news contents into morpheme, and to tag positive/negative polarity of each word by comparing this with sentimental dictionary. Third, positive/negative polarity of news is judged by using indexed classification information and scoring rule, and then final investment decision making information is derived according to daily scoring criteria.

For this study, KOSPI index and its fluctuation range has been collected for 63 days that stock market was open during 3 months from July 2011 to September in Korea Exchange, and news data was collected by parsing 766 articles of economic news media M company on web page among

* Graduate School of Business IT, Kookmin University
** School of MIS, Kookmin University
*** Corresponding Author : Seung Ryul Jeong
School of Management Information Systems, Kookmin University 861-1, Jeongneung-dong, Seongbuk-gu, Seoul 136-702, Korea.
Tel: +82-2-910-4568, Fax: +82-2-910-5209, E-mail: srjeong@kookmin.ac.kr

article carried on stock information>news>main news of portal site Naver.com. In change of the price index of stocks during 3 months, it rose on 33 days and fell on 30 days, and news contents included 197 news articles before opening of stock market, 385 news articles during the session, 184 news articles after closing of market.

Results of mining of collected news contents and of comparison with stock price showed that positive/negative opinion of news contents had significant relation with stock price, and change of the price index of stocks could be better explained in case of applying news opinion by deriving in positive/negative ratio instead of judging between simplified positive and negative opinion. And in order to check whether news had an effect on fluctuation of stock price, or at least went ahead of fluctuation of stock price, in the results that change of stock price was compared only with news happening before opening of stock market, it was verified to be statistically significant as well.

In addition, because news contained various type and information such as social, economic, and overseas news, and corporate earnings, the present condition of type of industry, market outlook, the present condition of market and so on, it was expected that influence on stock market or significance of the relation would be different according to the type of news, and therefore each type of news was compared with fluctuation of stock price, and the results showed that market condition, outlook, and overseas news was the most useful to explain fluctuation of news. On the contrary, news about individual company was not statistically significant, but opinion mining value showed tendency opposite to stock price, and the reason can be thought to be the appearance of promotional and planned news for preventing stock price from falling.

Finally, multiple regression analysis and logistic regression analysis was carried out in order to derive function of investment decision making on the basis of relation between positive/negative opinion of news and stock price, and the results showed that regression equation using variable of market conditions, outlook, and overseas news before opening of stock market was statistically significant, and classification accuracy of logistic regression accuracy results was shown to be 70.0% in rise of stock price, 78.8% in fall of stock price, and 74.6% on average.

This study first analyzed relation between news and stock price through analyzing and quantifying sensitivity of atypical news contents by using opinion mining among big data analysis techniques, and furthermore, proposed and verified smart investment decision making model that could systematically carry out opinion mining and derive and support investment information. This shows that news can be used as variable to predict the price index of stocks for investment, and it is expected the model can be used as real investment support system if it is implemented as system and verified in the future.

Key Words : Big Data, News contents, Opinion Mining, Text Mining, Sentimental Analysis, Stock Index, Stock Investment

저 자 소 개

김유신

국민대학교 정보관리학과에서 학사, 비즈니스IT전문대학원에서 석사학위 취득 후 박사과정에 재학 중이며, 현재 SKC&C 금융사업본부에 재직 중이다. 금융과 의료분야에서 다수의 웹・모바일 프로젝트를 수행하였으며, 인터넷뱅킹, 보험사이버창구, 퇴직연금, VOC, Refer 시스템 등을 구축하였다. 주요 관심분야는 오픈웹/웹접근성, SNS, CRM, 오피니언 마이닝, 빅데이터 분석 등이다.

김남규

현재 국민대학교 경영정보학부에서 조교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 편집위원, 한국지능정보시스템학회 및 한국 CRM학회 이사, 그리고 한국경영정보학회 및 한국정보시스템학회 종신회원으로 활동 중이다. 주요 관심분야는 시멘틱 데이터 관리, 데이터베이스 설계 및 데이터마이닝 등이다.

정승렬

서강대학교에서 경제학사, 미국 위스컨신 대학에서 경영정보학 석사, 그리고 사우스케롤라이나 대학에서 경영정보학 박사를 취득하였다. 현재 국민대학교 경영정보학부 및 비즈니스IT전문대학원 교수로 재직중인 그는 Journal of MIS, Communications of the ACM, Information and Management, Information Systems Management, Journal of Systems and Software, Lecture Notes on Computer Science, APJIS, 경영과학, 한국경영과학회지, ISR, 정보처리학회지 등의 국내외 저널에 프로세스 관리, ERP, 정보자원관리, 시스템 구현, 정보시스템 감리 등의 주제와 관련하여 많은 논문을 발표하였다