JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 7 : 1593~1598 Jul. 2015

사회적 감성과 주가의 상관성 분석

윤홍원*

Correlation Analysis of Social Sentiment and Stock Prices

Hongwon Yun*

Department of Computer & Information Engineering, Silla University, Pusan 617-736, Korea

요 약

본 논문에서는 사회적 감성과 주가의 상관성을 분석한다. 먼저, 주가 폭락 또는 폭등 기간과 그 직전의 극성을 각각 분석하고 이 결과를 이용하여 사회적 감성과 주가 사이의 상관관계를 분석한다. 본 연구를 위하여 과거의 다우존스산업평균지수 데이터를 수집하고 주가의 폭등과 폭락 시점을 검출한다. 검출한 시점에 근거하여 뉴욕 타임즈 기사를 수집하고 극성을 분석한다. 분석 결과에 의하면 주가 폭락 기간보다 폭등 기간에는 부정적 용어의 출현 빈도가 감소하고 긍정적 용어의 출현 빈도가 증가한다. 주가 폭락 또는 폭등 직전에는 부정적 용어의 출현 빈도와 긍정적 용어의 출현 빈도 사이에 차이가 커지 않다. 상관관계 분석에 의하면, 주가 폭락과 폭등 기간에는 사회적 감성과 주가 사이에 양의 상관관계를 보인다. 반면에, 주가 폭락과 폭등 직전에는 사회적 감성과 주가 사이에 유의한 수준의 상관관계를 나타내지 않는다.

ABSTRACT

In this paper, we analyze the correlation between social sentiment and stock prices. Polarity analysis is conducted for the stock prices plunging and soaring duration. And it is performed for its prior period. Using these results, we analyze the relationship between the social sentiment and stock prices. We collected the past data of Dow Jones Industrial Average and detected the period of plunging and soaring. On the basis of the detected time, the New York Times articles are collected and polarity analysis is conducted. Frequency of negative terms is decreased and it of positive terms is increased during the stock prices soaring. There is a little difference between the frequency of negative and positive terms in the previous stock prices plunging or soaring. According to the correlation analysis, it shows a positive correlation between social sentiment and stock prices in the period of plunging and soaring. A significant correlation is not appeared in the previous stock prices plunging or soaring.

키워드 : 감성 분석, 극성 분석, 빅 데이터, 텍스트 분석

Key word : Sentiment analysis, Polarity analysis, Big data, Textual analysis

Received 08 June 2015, **Revised** 22 June 2015, **Accepted** 06 July 2015

*** Corresponding Author** Hongwon Yun(E-mail:hwyun@silla.ac.kr, Tel:+82-51-999-5065)
Department of Computer & Information Engineering, Silla University, Pusan 617-736, Korea

Open Access **http://dx.doi.org/10.6109/jkiice.2015.19.7.1593** print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

지금까지 주가에 영향력 있는 요소에 관하여 투자자와 연구자들은 많은 관심을 가지고 연구를 진행하고 있다[1-3]. 미디어에서 제공하는 뉴스와 같은 자료에서부터 트위터(Twitter) 메시지와 같은 SNS(Social Networking Service)에 데이터까지 다양한 자료를 주가 예측 자료로 활용하고 있다[4-6]. 주가에 영향을 미치는 요소가 다양하고 가중치는 가변적이므로 주가 예측은 여전히 산업계와 학계에 관심 있는 주제로 남아 있다[7-9]. 높은 신뢰 수준의 주가 예측은 여전히 어려운 문제이지만 주가가 폭락하거나 폭등하는 시기의 사회적 분위기를 계량화 할 수 있다면 증시의 흐름을 파악하는데 도움을 받을 수 있다[10-12]. 또한 주가가 폭락이나 폭등하기 직전의 사회적 분위기가 온라인 미디어에 잠재해 있는지 빅 데이터를 활용한 실증적인 연구가 필요하다.

본 연구에서는 주가 폭락이나 폭등 시기의 사회적 감성을 분석하기 위하여 과거의 주가 데이터를 수집하고 주가가 폭락한 시기와 폭등한 시기를 찾아낸다. 주가 폭락이나 폭등 시기의 사회적 분위기가 잠재할 가능성이 있는 자료로써 온라인 뉴스 기사를 수집한다. 감성 분석 도구를 사용하여 수집한 뉴스 기사의 극성(polarity)을 분석하고 이 결과를 바탕으로 주가가 폭락하거나 폭등하기 직전의 사회적 분위기를 분석한다. 또한 실제 주가가 폭락 또는 폭등한 시점의 사회적 분위기를 분석한다. 본 연구를 통하여 주가 폭락이나 폭등에 앞서 사회적 감정이 부정적이거나 긍정적으로 잠재하는지 알아보고 실제로 주가 폭락이나 폭등이 일어난 시점의 사회적 감정은 어떻게 나타나는지 분석한다.

본 논문의 구성은 다음과 같다. 1장에서는 기존 연구의 동향과 본 연구의 배경을 기술하였다. 본 연구에 필요한 데이터의 수집 방법과 분석 방법은 2장에서 기술한다. 3장에서는 분석한 결과를 나타내고 의미를 해석한다. 마지막으로, 4장에서는 결론을 맺고 남은 연구 과제를 알아본다.

Ⅱ. 데이터 수집 및 분석 방법

2.1. 데이터 수집 및 예비 분석

먼저 주가의 폭락과 폭등 시기를 찾기 위하여 야후 금융 사이트(http://finance.yahoo.com)에서 과거 데이터를 수집하였다. 이 사이트에서 2008년 1월 2일부터 2013년 12월 31일까지 6년 동안의 다우존스산업평균지수(Dow Jones Industrial Average, 이하 DJIA) 데이터를 1521개 수집하였다. 그림1은 2008년 1월부터 2013년 12월까지 DJIA 의 흐름을 보이고 그림 2는 이 기간 동안 수집한 DJIA 데이터의 샘플을 나타낸다. 수집한 데이터는 날짜, 시가, 최고가, 최저가, 종가, 거래량을 포함하고 있다.

수집한 6년간의 과거 DJIA 데이터에 대하여 연도별로 5일 간격으로 연속적인 기울기를 구하였고 그 결과는 그림 3과 같다. 기울기가 큰 날짜부터 정렬하여 주가가 폭락한 시기와 폭등한 시기를 찾았다. 주가 폭락 시기와 폭등 시기의 날짜 수가 유사하게 겹쳐지는 기울기의 절대값이 200이었으며 이 값을 기준으로 폭락과 폭등한 시기의 값을 추출하였다.

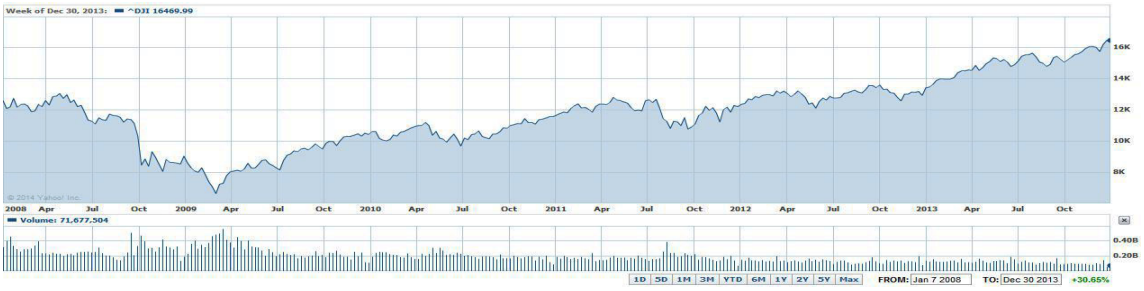

그림 1. DJIA의 과거 주가 흐름도 (2008.1~2013.12)
Fig. 1 DJIA Chart for historical stock prices from Jan. 2008 to Dec. 2013

Date	Open	High	Low	Close	Volume
02-Jan-08	13,261.82	13,279.54	12,991.37	13,043.96	239,580,000
03-Jan-08	13,044.12	13,137.93	13,023.56	13,056.72	200,620,000
04-Jan-08	13,046.56	13,046.72	12,789.04	12,800.18	304,210,000
07-Jan-08	12,801.15	12,884.15	12,733.84	12,827.49	306,700,000
08-Jan-08	12,820.90	12,906.42	12,565.41	12,589.07	322,690,000

그림 2. 수집한 DJIA 데이터의 샘플
Fig. 2 Collected sample DJIA data

실제 주가 폭락이 시작한 날짜로부터 1일에서 5일전까지를 주가 예측을 위한 기간으로 설정하였다. 마찬가지로 폭등 예측 기간은 실제 폭락 5일전까지로 설정하였다. 예측을 위한 사전 기간과 실제 기간을 그림으로 나타내면 그림 4와 같다. 앞으로 '실제 기간'은 주가가 실제로 폭락하였거나 폭등한 기간을 뜻하고, '사전 기간'은 주가가 폭락하거나 폭등하기 시작한 날짜로부터 1일에서 5일전까지를 뜻한다. 뉴스 기사를 수집하기 위하여 미국의 주요 일간지인 뉴욕 타임즈(The New York Times)를 선정하였다. 실제 기간과 사전 기간에 맞추어 뉴욕 타임즈의 과거 기사 모음 사이트에서 총 9,000개의 온라인 기사를 수집하였다.

2.2. 극성 분석 방법

실제 기간과 사전 기간에 맞추어 수집한 뉴욕 타임즈의 온라인 기사를 분석하기 위하여 감성 분석(sentiment analysis)을 시도하였다. 본 연구에서는, 피츠버그 대학교, 코넬 대학교, 유타 대학교의 연구진들이 공동 개발하였고 인터넷에 공개되어 있는 오피니언파인더 버전2.0 (OpinionFinder 2.0)을 극성 분석 도구로 사용하였다.

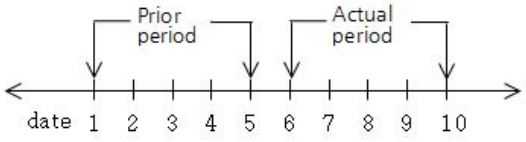

그림 4. 사전 기간과 실제 기간
Fig. 4 Prior period and actual period

오피니언파인더는 문서를 읽어서 극성을 분석할 수 있는데 "긍정적(positive)", "부정적(negative)", "중립적(neutral)"으로 분류한다. 오피니언파인더는 우리가 수집한 온라인 뉴스 기사를 입력 데이터로 하고 긍정적 용어 개수, 부정적 용어 개수 그리고 중립적 용어 개수를 결과로 출력한다.

Ⅲ. 분석 결과

앞의 2장에서 언급한대로 과거 6년 동안의 DJIA 데이터를 분석하여 주가 폭락과 폭등 기간을 검출하였다. 이 기간에 맞추어 뉴욕 타임즈의 기사를 9,000개 수집하였고 이것을 오피니언파인더로 극성 분석을 하였더니 표 1과 같은 결과가 나왔다.

사회적 감성 지수는 식(1)과 식(2)와 같이 정의한다. 주가의 폭락 기간을 d 라고 하고 사회적 감성 지수를 SIL이라고 하면 폭락 기간의 사회적 감성 지수는 다음과 같이 정의한다.

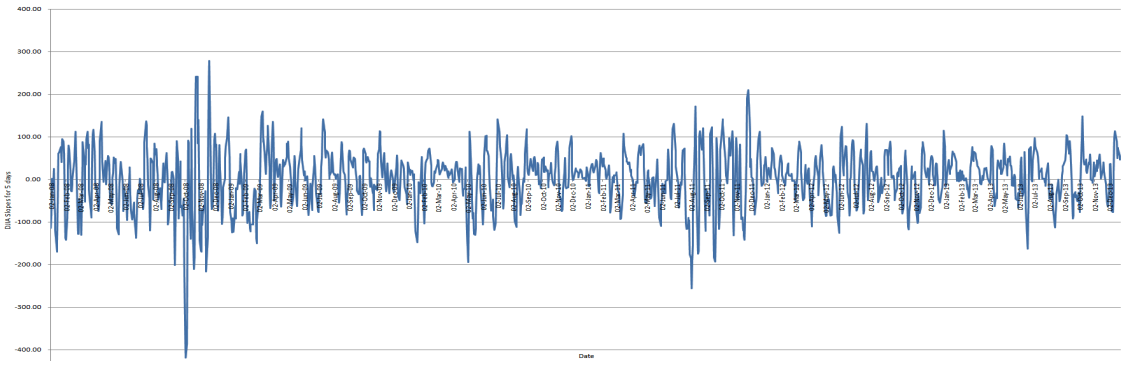

그림 3. 6년간의 DJIA 데이터의 연속적인 기울기
Fig. 3 DJIA Slopes for six years

표 1. 극성 분석의 결과
Table. 1 Result of polarity analysis

	Sum of terms during plunging			Sum of terms during Soaring		
	Neutral	Negative	Positive	Neutral	Negative	Positive
Prior period	5034	587	279	5107	663	390
Actual period	5210	579	224	4988	612	388

단, n_i는 폭락 기간에 어떤 날짜의 부정적 용어의 출현빈도를 나타내고 p_i 는 폭락 기간 중 어떤 날의 긍정적 용어의 출현빈도를 나타낸다.

$$SIL_d = \frac{\sum_{j=1}^{n} n_j}{\sum_{i=1}^{m} (n_i + p_i)} \quad (1)$$

폭등 기간의 사회적 감성 지수는 아래 식과 같이 SIO로 정의한다.

$$SIO_d = \frac{\sum_{j=1}^{n} p_j}{\sum_{i=1}^{m} (n_i + p_i)} \quad (2)$$

위 식을 사용하여 사회적 감성 지수를 계산하였다. 그림 5는 주가 폭락과 폭등 기간의 사회적 감성을 나타내고 있다. 이 그림에 의하면 주가 폭락 기간보다 폭등 기간에는 부정적 용어의 출현 빈도가 감소하고 긍정적 용어의 사용이 증가하였음을 알 수 있다. 그림 6은 주가 폭락과 폭등의 사전 기간에 나타난 부정적 용어와 긍정적 용어의 출현빈도를 보인다. 폭락 기간의 부정적 용어 출현 빈도와 폭등 기간의 부정적 용어 출현 빈도를 비교하면 약간의 차이가 있는데 부정적 용어의 출현 빈도가 감소하였다. 긍정적 용어의 출현 빈도는 조금 높음을 볼 수 있다. 그림 5와 그림 6을 비교하면 실제 기간이 사전 기간보다 사회적 감성을 더 많이 드러내고 있다. 이 차이를 추세선으로 나타내면 그림7과 같다. 그림 7은 주가의 변화에 따른 긍정적 용어의 추세를 보이고 있는데 실제 기간이 사전 기간보다 기울기가 큼을 알 수 있다. 이것은 실제 기간이 사전 기간보다 사회적 감성을 더 많이 드러내는 경향이 있음을 나타낸다.

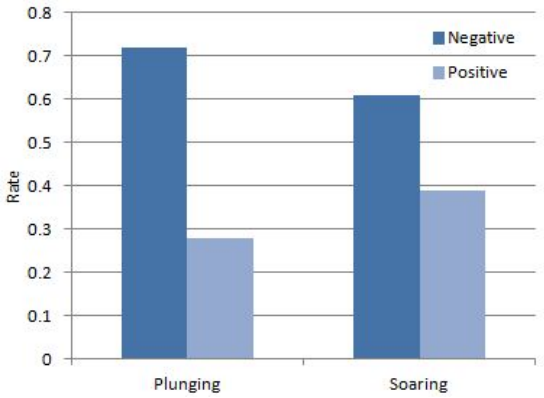

그림 5. 주가 폭락과 폭등 실제 기간에서 사회적 감성 비교
Fig. 5 Comparison of social sentiment between actual plunging and soaring duration

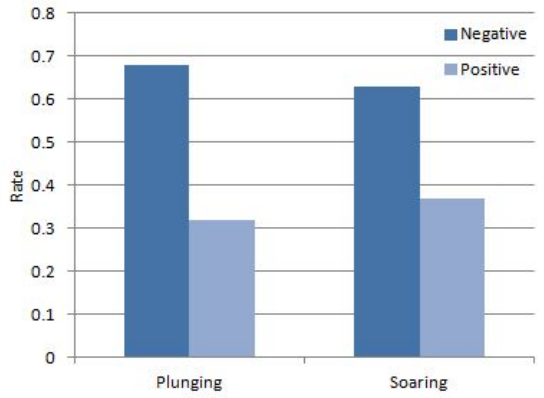

그림 6. 주가 폭락과 폭등 사전 기간에서 사회적 감성 비교
Fig. 6 Comparison of social sentiment between prior plunging and soaring duration

폭락 기간과 폭등 기간 사이의 주가변화에 따른 부정적 용어와 긍정적 용어 사이의 상관관계를 분석하였다. 아래와 같은 피어슨 상관관계 분석을 이용하여 상관계수를 구하였다.

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \quad (3)$$

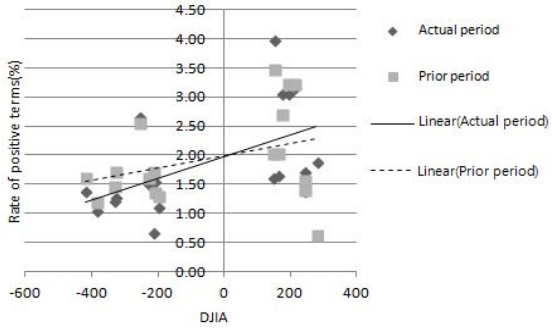

그림 7. 실제 기간과 사전 기간에서 사회적 감성의 추세선 비교
Fig. 7 Comparison of social sentiment trend lines between actual period and prior period

표 2. 사회적 감성과 주가의 상관계수
Table. 2 Correlation coefficient of social sentiment and stock prices

	Negative	Positive
Prior period	0.33	0.34
Actual period	0.20	0.53

표 2는 사전 기간과 실제 기간 사이에 부정적 용어와 긍정적 용어의 출현에 관한 상관관계를 나타낸다. 사전 기간에는 부정적 용어와 긍정적 용어 사이에 상관관계의 차이가 적지만 실제 기간에는 긍정적 용어가 비교적 강한 양의 상관관계를 보인다.

Ⅳ. 결 론

본 연구에서는 주가 폭락과 폭등 시기에 사회적 감성을 분석하였다. 또한 주가 폭락이나 폭등 시기의 직전에 사회적 감성이 사전에 잠재하는지 분석하였다. 본 연구를 위하여 과거 6년 동안의 다우존스산업평균지수 데이터를 수집하고 주가의 폭락과 폭등 시점을 먼저 검출하였고 사전 분석을 하였다. 이 사전 분석 결과를 바탕으로 실제 폭락과 폭등 기간을 정하고 폭락과 폭등의 사전 기간을 설정하였다. 실제 폭락과 폭등 기간과 사전 기간에 맞추어 뉴욕 타임즈의 기사를 수집하고 극성 분석을 시행하였다. 극성 분석의 결과에 의하면, 실제 폭락 기간보다 폭등 기간에는 부정적 용어의 출현 빈도가 감소하고 긍정적 용어의 출현 빈도가 증가함을 알 수 있었다. 사전 기간에는 폭락 기간과 폭등 기간 사이에 부정적 용어와 긍정적 용어의 출현 빈도에 약간의 차이를 보였다. 주가의 변화에 따른 긍정적 용어의 출현 추세에서는 실제 기간이 사전 기간보다 좀 더 명확한 추세를 보였다. 사전 기간과 실제 기간 사이에 부정적 용어와 긍정적 용어의 출현에 관한 상관관계 분석에 따르면 실제 기간에는 주가 상승과 긍정적 용어 사이에 양의 상관관계가 나타났으며 사전 기간에는 유의한 차이를 보이지 않았다. 주가 관련 데이터의 분석에서 실제 기간이 사전 기간보다 사회적 감성을 분명하게 나타냄을 알 수 있다.

REFERENCES

[1] T Kimoto, K Asakawa, M Yoda and M Takeoka, "Stock market prediction system with modular neural networks," *1990 International Joint Conference on Neural Networks,* June 17-21, San Diego, CA, USA, Vol. 1, pp. 1-6, 1990.

[2] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news," *ACM Transactions on Information System*, Vol. 27, Issue 2, pp. 1-19, 2009.

[3] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah and D. C. LingNgo, "Text mining for market prediction: A system review," *Expert Systems with Applications*, Vol. 41, No. 16, pp. 7653-7670, 2014.

[4] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, Vol. 2, Issue 1, pp. 1-8, 2011.

[5] Y. Kim, N. Kim, S. R. Jeong, “Stock-index Invest Model Using News Big Data Opinion Mining,” *Journal of Intelligence and Information Systems*, Vol. 18, No. 2, pp.143-156, 2012.

[6] A. Porchnev, N. Novgorod, I. Redkin and A. Shevchenko, "Stock market indicator based on historical data and data from twitter sentiment analysis," *2013 IEEE 13th International Conference on Data Mining Workshops*, Dallas, TX, USA, pp. 440-444, 2013.

[7] B. Li, K. C. C. Chan and C. Ou, "Public sentiment analysis in twitter data for prediction of a company's stock price movement," *2014 IEEE 11th International Conference on e-Business Engineering*, Guangzhou, China, pp. 232-239, 2014.

[8] R. G. Lin and T. Tsai, "Scalable system for textual analysis of stock market prediction," *The 3th International Conference on Data Analysis*, Rome, Italy, pp. 95-99, 2014.

[9] A. Gupta and S. D. Sharma, "Clustering-Classification based prediction of stock market future prediction," *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 3, pp. 2806-2809, 2014.

[10] I. Markovic, M. Stojanovic, M. Bozic and J. Stankovic, "Stock market trend prediction based on the LS-SVM model update algorithm," *Advances in Intelligent Systems and Computing*, Vol. 311, pp. 105-114, 2015.

[11] J. Patel, S. Shah, P. Thakkar and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert Systems with Applications*, Vol. 42, Issue 4, pp. 2162-2172, 2015.

[12] R. A. Araujo, A. Oliveria and S. Merita, "A hybrid model for high-frequency stock market forecasting," *Expert Systems with Applications*, Vol. 42, Issue 8, pp. 4081-4096, 2015.

윤홍원(Hongwon Yun)

1996년 ~ 현재 신라대학교 컴퓨터정보공학부 교수
1998년 부산대학교 전자계산학과 박사
2003년 NCSU 교환교수
2010년 TAMU 교환교수
※관심분야 : 데이터베이스, 빅데이터