텍스트마이닝과 환율데이터를 활용한 주가의 위험성 예측에 관한 연구

A Study on the Prediction of Stock Market Crisis by Text Mining and Exchange Rates

변태우, 이성우, 김재광, 이지형
Tae-u Byeon, Sung-Woo Lee, JaeKwang Kim, Jee-Hyong Lee

성균관대학교 정보통신공학부
E-mail: {specialbtw, lsmoney, linux}@skku.edu, jhlee@ece.skku.ac.kr

요 약

최근 우리나라 주가-환율 상관관계가 높아진 점에 착안하여, 본 논문에서는 전통적인 주가예측기법인 뉴스 텍스트마이닝 기법에 환율데이터라는 변수를 추가로 접목하여, 향후 주식시장의 장·단기적 침체기를 효율적으로 예측할 수 있는 방법을 제시한다. 먼저, 본 논문에서는 시장의 불안정성을 예측하기 위한 위험예측모델을 정의한다. 제안된 모델에서는 특정단어의 빈도수가 일정 수 이상 높을 경우, 주가하락이 발생한다고 간주하며, 이와 동시에 환율이 일정비율 이상 상승할 경우 주가가 하락한다가 간주한다. 본 연구는 포털사이트 네이버API를 이용하여 수집된 기사와 한국거래소에서 추출한 KOSPI지수 데이터 등을 대상으로 실험을 수행하였다.

키워드 : 주식, 텍스트마이닝, 환율, 주가예측, 위험성

1. 서 론

일반적으로 주식가격을 예측하는 것은 주식가격변동에 영향을 미치는 많은 요인간의 상호작용으로 인해 매우 어렵다고 알려져 있다. 주식시장 정보는 주어지는 자료가 방대하고, 잡음이 포함되어 있는 경우가 많아 예측의 불확실성이 높은 특징을 가지고 있다. 그럼에도 불구하고 현재에 이르기까지 경제, 수학, 통계 및 전산 분야에 걸쳐 오랜 기간 동안 주가예측을 위한 다양한 연구가 시도되고 있으며, 그 결과, 신경망, SVM, 유전자 알고리즘, 역전파 알고리즘 등 다양한 주가예측기법이 개발되었다.

본 연구에서는 뉴스 텍스트 마이닝 및 환율데이터 분석을 바탕으로 주가하락의 위험구간을 조기 예측·판단하고자 한다. 정확한 예측을 위해 2001년 8월부터 2010년 12월까지 총 10여년간의 경제뉴스를 표본으로 텍스트마이닝을 수행하였으며, 특정단어의 빈도수가 주가변동에 미치는 영향에 대한 학습을 실시하고, 이에 환율이라는 거시경제 요소의 등락을 추가변수로 구성하여, 종합주가지수(KOSPI) 변동의 위험구간을 예측하는 모형을 제안한다.

근래에 들어서 뉴스데이터로부터 주가를 예측하고자 하는 연구가 활발히 이뤄지고 있다. 뉴스와 주가간의 강력한 상관관계는 비교적 최근의 연구 결과에서도 증명된 바 있다. 금융 뉴스를 이용한 주가 예측기법은 대개 다음과 같은 접근법을 취한다. 뉴스데이터를 수집한 이후 이 뉴스데이터에 대해서 텍스트 마이닝 처리를 하여 문서내의 의미 있는 특징들을 추출한 후, 이를 이용하여 해당 뉴스가 주가에 호재인지, 악재인지를 분류한다. 그리고 분류된 결과를 이용하여 시뮬레이션 투자 및 가격변동추이 예측을 시도한다. 뉴스 본문의 자연어 처리 방법으로는 bag of words, TF-IDF방식이 주로 사용 되었으며, 유사한 의미의 단어들을 묶어서 의미를 확장하는 명사구 처리기법과 유사한 logical AND 기법이 사용된 예도 있다. 또한 자연어 문서처리에 있어서 가장 간단한 방법인 PR 역시 최근 연구에서 사용된 사례가 있다. 그 외에 기존에 널리 사용되는 bag of words 방법보다 고유명사 처리기법을 이용하여 보다 의미있는 결과를 얻었다는 보고도 있다. 추출된 특징으로부터 해당 뉴스를 분류하는 방법으로는 naïve Bayesian 분류기와 SVM이 많이 사용되고 있다.

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업의 연구 결과입니다. 연구비 지원에 감사드립니다(No. 2009-0075109).

2. 텍스트 마이닝

텍스트마이닝이란 비정형 텍스트 데이터에서 가치와 의미가 있는 정보를 찾아내는 기술을 말한다. 사용자는 텍스트마이닝 기술을 통해 방대한 정보뭉치에서 의미있는 정보를 추출해내고, 다른 정보와의 연계성을 파악하며, 텍스트의 카테고리를 찾아내는 등, 단순한 정보검색 그이상의 결과를 얻어낼 수 있다. 텍스트마이닝에서 다루고 있는 기술 분야는 문서분류(Document Classification), 문서군집(Document Clustering), 정보추출(Information Extraction), 문서요약(Document Summarization) 등을 들 수 있다. 본 논문에서는 정보추출 기술을 응용하여 데이터를 수집·가공하고, 다음과 같이 실험하였다.

1) 질의어는 (“주식”, “상승”, “매수”, “사라”) 등의 긍정적 뉴스 검색에 유리한 키워드로 한다.
2) 검색된 문서의 수는 월 단위로 기록한다.

<표 1> 추출된 표본 데이터

처리되는 주가간격	데이터 수집기간	학습데이터 기간	대상 주식시장
거래일 종가	2002-2010	60일	KOSPI

3. 환 율

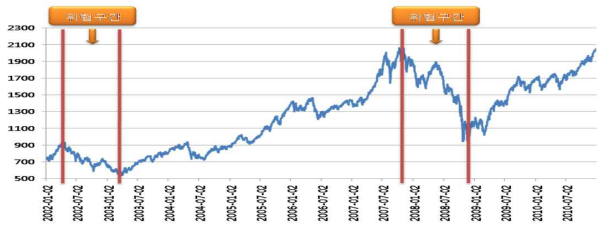
환율이란 자국통화와 외국통화의 교환비율을 말한다. 일반적으로 환율이 변하게 되면 국내의 수출주도형 기업들이 영향을 받게 되므로, 주가에 상당한 영향을 미치는 것으로 알려져 있다. 본 논문에서는 환율의 변동성을 분석하여 주가에 미치는 영향을 미리 예측한다.

4. 텍스트마이닝-환율데이터를 이용한 주가 위험성 예측

<그림 1> 위험예측모델의 구성

본 연구에서 제안한 위험예측시스템은 Training 단계와 Testing단계로 구성되며, 경제위기가 발생하여 시장이 그것을 인지하여 반응하고 있는 구간을 위험구간이라 정의한다. Training단계에서는 2001년 8월부터 2010년 12월 사이에서 위험구간을 찾아 특징을 추출하고, 추출된 특징을 이용하여 패턴을 학습한 뒤, 학습된 data를 Traing DB에 저장한다.

Testing 단계에서는 텍스트마이닝으로 얻어낸 뉴스특정단어빈도수와 환율데이터를 분석하는 패턴 분류과정을 거친다. 패턴인식의 결과로 위험구간으로 인식될 경우 알림을 발생시킨다.



<그림 2> KOSPI 주가지수의 위험 구간 설정

한국거래소(http://www.krx.co.kr)에서 얻은 주가데이터를 분석하여, 주가지수가 큰 폭으로 하락한 2002년 7월9일부터 8월일까지의 기간과 2007년 10월 11일부터 2009년 1월 15일까지의 기간을 위험구간으로 설정하였다.

<표 2> 위험구간 내의 패턴인식률

항목	범위	대상수	예측 성공	예측 실패	예측 성공률
단어빈도수 & 환율	N>= 4	10	6	4	60%
	r>=1%				

4.1 뉴스 속 특정단어 빈도수 - 주가 관계 분석

특정단어 빈도수를 확인하기 위해 질의어는 “주식”, “상승”, “매수”, “사라” 등으로 지정 하였으며, 이 질의어들이 발생한 후 일정시간 뒤 주가 하락이 발생한다고 가정하였다.

4.2 환율 - 주가 관계 분석

또 다른 판단기준인 환율은 원-달러환율을 선택하였고, 환율이 일정비율 이상 상승할 경우(원화가치가 하락) 주가가 하락한다는 사실을 가정하였다.

1) 뉴스의 단어빈도수 체크를 위한 질의어는 (“주식”, “상승”, “매수”, “사라”) 등으로 한다.
2) 검색된 문서의 수는 월단위로 기록한다.
3) 전월 환율의 증감으로 익월 주가의 흐름을 예상한다.
4) 실험의 예측성공률을 높이기 위해 단어빈도수 및 환율 범위를 보정한다. (단어빈도수>=10, 환율>=4%)

<표 3> 실험데이터의 예측성공률

항목	조건 범위	해당 대상수	예측 성공	예측 실패	미결정	예측 성공률
단어 빈도수	N>=10	36	12	22	2	33.3%
	N< 10	70	27	39	4	38.5%
환율	r>=4%	8	3	5	0	37.5%
	r< 4%	98	36	56	6	36.7%
단어빈도수 & 환율	N>= 10	4	2	2	0	50%
	r>=4%					

5. 실험 결과 및 결론

본 연구에서는 뉴스 데이터마이닝과 환율데이터를 이용해 주가에 미치는 영향을 연구해보았다. 단어빈도수와 환율을 각각 활용한 예측에서는 최대 38.5%와 37.5%의 다소 낮은 예측성공률을 보였으나, 단어빈도수와 환율을 함께 예측에 활용한 결과 예측성공률이 50%로 높아지는 성능의 향상을 보였다. 뉴스를 이용한 기존의 주가 예측 방법은 뉴스 자체의 정보만을 이용하여 주가를 예측할 수 있는지에 대해 주로 연구 되었다. 더 높은 예측성공률을 위해 향후에는 환율 이외에 물가지수, 금리 등을 활용한 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] J.D.Thomas, K.Sykara, “Intergraing Genetic Algorithms and Text Learning for Financial Prediction”, in Proceedings of the Genetic and Evolutionary Computing 2000 Conference Workshop on Data Mining with Evolutionary Algorithms, pp. 72-75, 2000
[2] S.W.Ahn, S.B.Cho, “Stock prediction using news text mining and time series analysis”, Korea Computer Congress, Vol 37, No.1(C), pp. 364-369, 2010
[3] B.S.Cho, “An Empirical Study on the Impact of News in KOSPI : The case of big 10 news announcements by broadcating”, 2009
[4] G.Gidofalvi, “Using News Articles to Predict Stock Price Movement”, Department of Computer Science and Engineering University of California, 2001
[5] R.P.Schumaker, H.Chen, “Textual analysis of stock prediction using breaking financial news : The AZFin text system”, ACM Transactions on Information Systems Volume 27 Issue 2, 2009
[6] S.G.Kim, "Empiricial Evidence for the Relationship between foreign Exchange Rate and Stock Price", Pusna National University, 2010