주식 관련 기사 분류 및 긍정 부정 판단을 통한 종목 추천 시스템

이유준O, 박정우, 전민재, 최준수, 한광수
국민대학교 컴퓨터공학부
yj_lee@kookmin.ac.kr, jungwooc@kookmin.ac.kr, mj_jeon@kookmin.ac.kr,
jschoi@kookmin.ac.kr, kshahn@kookmin.ac.kr

Stocks Recommending System through Classifying News Articles by Positive or Negative Decision

Yoojun LeeO, Jungwoo Park, Minjae Jeon, Joonsoo Choi, Kwangsoo Hahn
Kookmin University

요 약

주식 시장에서 거래되고 있는 증권은 MACD(Moving Average Convergence Divergence), Stochastic 등의 보조 지표를 이용하는 기술적 분석을 통하여 매수/매도 시점을 결정한다. 주식 시장의 객관적인 자료를 통하여 분석하는 기술적 분석 방법은 주식 시장 외적인 요소를 반영하는데 있어 한계점이 존재한다. 본 논문에서는 기술적 분석 방법에 기사를 종목별로 분류하고 기사의 긍정 및 부정을 판별하는 문서 분류 기법을 적용하여 주식 외적인 요소를 반영하는 시스템을 제안한다.

주제어: Genetic algorithm, Stocks recommendation, Opinion mining, Document classification

1. 서론

주식 투자에 있어 기술적 분석은 과거의 주가와 거래량을 이용해 주식의 추세를 분석하여 미래의 주가를 예측하는 방법이다[1]. 그러나 기술적 분석을 통한 방법은 완벽히 신뢰할 수 있는 방법은 아니다. 가령 특정 회사에서 새로운 기술을 개발했거나 국가적 위기 상황과 같은 주식 시장 외적인 요소들은 기술적 분석 방법만으로 반영 할 수 없다. 최근 인터넷의 보급화로 인해 모든 소식들을 인터넷 기사로 접할 수 있게 되었다. 이러한 기사들은 신문사나 기자에 따라 주관적인 특성을 가지고 있다. 이런 기사들은 주식 시장에 직접적 혹은 간접적으로 영향을 준다.

현대 세계에서는 엄청난 양의 정보들을 생성하고 있다. 응용 분야의 광범위한 범위에서 방대한 규모의 데이터들이 계속해서 수집되고 있는데, IBM에서는 매일 2.5 퀸틸리언 바이트(2.5 quintillion byte, 2.5×10^{18} byte)의 데이터가 발생함을 언급하고 있다[2]. 이처럼 인터넷 공간에는 사람이 처리할 수 없을 정도로 많은 기사들이 존재한다. 이와 같이 대용량 데이터 처리의 중요성이 부각되면서 대용량 데이터의 효율적인 저장, 관리, 검색, 분석 및 활용을 위한 연구가 활발히 이루어지고 있는 실정이다.

본 논문에서는 뉴스 기사에서 관심있는 종목에 대한 기사를 선별하여 해당 기사가 주가에 미치는 것을 주식 추천 시스템에 반영하고자 한다. 기술적 분석 방법을 통해 객관적인 데이터를 기반으로 하여 추천 종목의 신뢰성을 높이고, 기사들을 각 종목별로 분류하여 기사의 긍정 및 부정을 판단하는 방법을 통해 주관적인 데이터를 확보하여 주식 시장 외적인 요소를 반영하는 증권의 매수/매도 종목 추천 시스템을 제안한다.

2. 관련 연구

2.1 기술적 분석과 유전 알고리즘

본 연구에서는 MACD 지표를 사용하여 기술적 분석을 한다. MACD 지표를 이용하여 유전 알고리즘을 구성하는 방법과 이를 통해 종목을 추천하는 방향을 제시한다.

2.1.1 기술적 분석(Technical analysis)

기술적 분석에서 많이 사용되고 있는 지표인 Slow stochastic과 MACD는 실무에서 많이 사용되고 있으며 신뢰도가 높은 편이다.

MACD는 단기 EMA(Exponential Moving Average, 지수이동평균)와 장기 EMA의 차이이다. 보편적으로 단기 EMA의 기간은 12일을 사용하고, 장기 EMA의 기간은 26일을 사용한다. MACD signal은 k(보편적으로 9일)일간의 EMA를 나타낸다. 또한 MACD oscillator는 MACD와 MACD signal과의 차이를 나타낸다.

MACD Line	단기 EMA - 장기 EMA
MACD Signal	k 기간의 EMA
MACD Oscillator	MACD - MACD Signal

표 1 MACD 계산 공식

본 연구에서는 MACD 지표를 계산하는데 사용되는 기간을 보편적으로 사용하는 기간으로 고정하지 않고, 각 종목에 적합한 날짜를 직접 구하여 계산하고자 한다. MACD 지표에서는 단기, 장기, k 총 3개의 기간이 필요하다.

2.1.2 유전 알고리즘

유전 알고리즘은 진화의 원리를 문제 해결에 이용하는 방법론이다. 다윈의 진화론과 같이 다산, 생존 경쟁, 변이, 자연 선택, 진화의 과정을 반복하여 진행하게 된다. 유전 알고리즘은 NP-hard 문제에 있어 효과적이다. 하지만 문제 해결을 위한 효율적인 알고리즘이 존재하는 경우는 비효율적인 방법이 될 수도 있다. 보편적인 유전 알고리즘은 하나의 해를 염색체라고 표현하고 복수개의 염색체를 운영하여 최적의 해를 구한다. 일반적인 유전 알고리즘은 아래와 같은 방법으로 진행된다[3].

최초 난수를 발생시켜 n개의 유전자인 해집단을 생성하고, 해 집단의 적합도를 평가하게 된다. 적합도란 각각의 유전자를 이용해 MACD를 구하고 지표가 제공하는 매매시점을 이용해 특정 기간 동안 모의 투자를 하여 수익률을 계산한다. 적합도를 평가한 후 품질 비례 선택과 엘리트 보전 전략을 이용해 교배 할 두 개체를 선택하게 된다. 이렇게 선택 된 두 개체를 교차 연산과 변이 연산을 통해 새로운 두 개체를 생성하게 된다. 생성 된 개체가 기존 해집단의 개체보다 적합도가 높게 되면 대치 연산을 통해 개체를 대치하게 된다. 위와 같은 과정을 반복하여 최적의 해를 찾게 된다[4].

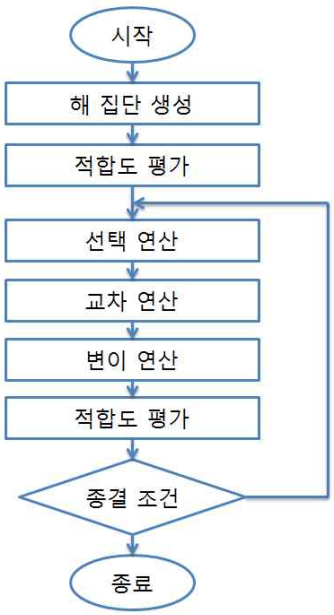

그림 1 유전 알고리즘 흐름도

단기 기간, 장기 기간, k 기간을 각각 8bit로 표현하여 하나의 24bit 유전자로 표현이 된다. 유전자를 이용하여 날짜를 변수로 한 최적의 MACD 지표를 구하게 된다.

x1 (단기)	1 <= x1 <= 256
x2 (장기)	1 <= x2 <= 256
x3 (k)	1 <= x3 <= 256

표 2 유전자 표기 방식

2.2 문서 분류

인터넷의 보급화와 사용 증가에 따라 웹상의 전자 문서의 양이 기하급수적으로 증가하고 있다. 전자 문서가 양적으로 크게 늘어남에 따라 이러한 수많은 문서를 사람이 분류하는 것은 거의 불가능해 졌다. 이에 따라 문서를 알맞게 분류하는 것을 도와주는 도구에 대한 요구가 점차 중요해 지고 있다.

문서 분류란 미리 정의되어 있는 두 개 이상의 범주(Category)에 대하여, 새로운 문서 집단이 입력되었을 때 미리 학습된 범주와 입력 문서 간의 유사도 비교를 통해 입력된 문서에 대한 범주를 결정해주는 기법이다[5]. 이러한 문서 분류는 다양한 분야에 적용되어 활용되고 있다. 스팸 메일 필터링, 뉴스기사 필터링, 뉴스기사의 주제 선정, 웹 문헌 분류 등 여러 분야에서 활발하게 연구되어지고 있다.

2.3 오피니언 마이닝

오피니언 마이닝(Opinion mining)은 감정 분석(Sentiment analysis)과 비슷한 의미로 사용된다. 텍스트에 포함된 내용이 주관적(Subjective)인지 객관적(Objective)인지 먼저 판별하고, 주관적이면 극성(Polarity)을 분석하여 내용이 긍정적(Positive)인지 부정적(Negative)인지 판별한다. 이러한 오피니언 마이닝은 대량의 정보 속에서 유용한 정보를 찾아내는 특징을 가지고 있다. 오피니언 마이닝은 특징(Feature) 추출, 의견 분류, 요약 및 표현 3가지 단계로 이루어진다.

특징 추출 단계에서는 중요한 정보로 판단되는 특징들을 추출해 낸다. 해당 특징의 어휘 정보도 추출된다. 의견 분류 단계에서는 추출된 특징과 어휘가 어떤 의미로 사용되었는지 판단 및 분류한다. 요약 및 표현 단계에서는 의견 성향이 밝혀진 의견정보들을 요약하여 전달하는 단계이다[6]. 이와 같은 오피니언 마이닝의 단계별 수행을 위해 다양한 방식으로 연구되고 있으며 자연어처리 기술 기반에서 통계적 기법에 이르기까지 여러 분야의 기술이 접목되고 있다. 기존에 긍정 및 부정을 판별하는 방법은 다음과 같다.

첫째로 SVMs(Support Vector Machines) 방법이 있다. 이는 미리 사전에 긍정 및 부정으로 분류된 학습 데이터(Training Sets)로 텍스트의 긍정 및 부정 의견을 분류하는 방식이다. 둘째로 N-grams or Part of Speech 방법이 있다. N-grams 단어 구조로 긍정 및 부정을 찾는 방식으로 문장을 Bi-gram decomposition으로 분리한다. 셋째로 Lexicon-based Approach 방법이 있다. 사전에 미리 정의된 긍정 및 부정 bag of words(1-grams or Uni-grams)를 이용하여 텍스트에 포함된 긍정 및 부정 단어의 출현 빈도로 긍정과 부정을 판별하는 기법이다. LIWC(Linguistic Inquiry and Word Count)나

POMS(Profile of Mood States) 같은 사전을 이용할 수 있는데 한국어에서는 사용할 수 없다. 이 방법에서는 긍정/부정 코퍼스(Corpus)를 만드는 것이 관건이다. 넷째로 Linguistic Approach 방법이 있다. 텍스트의 문법적인 구조를 파악하여 극성을 판별하는 기법이다. 주로 Lexicon-based Approach 방식과 함께 사용한다. 문맥(Context) 등을 파악하여 극성을 파악한다.

오피니언은 어떤 단위로 볼 것인가에 따라 3가지로 분류될 수 있다. 문서 전체적으로 의견을 종합하는 document level, 문서를 문장 단위로 나누어서 개별 문장의 의미를 파악하는 sentence level, 문장 내부의 어떤 개체의 속성에 대한 의견을 파악하는 entity and feature/aspect level로 분류 할 수 있다.

3. 기사 분류 및 분석 시스템

기술적 분석은 객관적인 주식 데이터를 기반으로 분석하기 때문에 주식 시장 외적인 요소들을 반영하는데 한계점이 있다. 하지만 실제 주식 시장에서는 해당 주식 종목에 대한 기사로 인해 주가에 영향을 미치게 된다. 본 논문에서 제안하는 기사 분석 시스템은 이러한 단점을 보완하여 안정적인 주식 종목 추천을 목표로 한다. 전체 시스템은 아래와 같다.

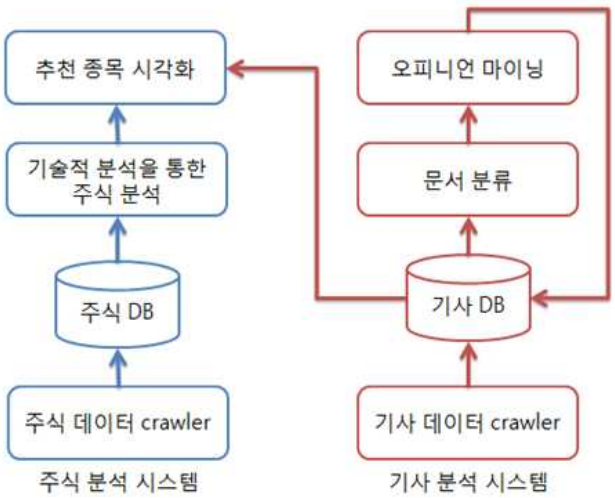

그림 2. 추천 종목 시스템의 구조

기사 분석 시스템은 실시간으로 신문 기사를 crawling 하고 이렇게 수집된 기사는 데이터베이스에 저장 되고, 문서 분류 컴포넌트에 입력된다. 주식 시장과 관련된 기사이기 때문에 다음과 같은 형식으로 문서가 분류된다. 기사 분류 시스템은 주식 종목당 2개의 카테고리로 분류가 된다. 하나는 주식 시장에 영향을 미치지 않는 기사 카테고리, 나머지 하나는 주식 시장에 영향을 미치는 기사 카테고리이다. 시스템 운용자가 처음 두 개의 카테고리 별로 기사를 학습시키고, 시스템에 입력되는 기사에 대해 분류를 하고 분류된 기사는 학습 된 데이터에 추가한다. 이렇게 분류되어 학습된 기사 데이터가 많이 쌓이게 되면 분류의 정확도는 높아지게 된다.

문서 분류 단계가 끝나면 오피니언 마이닝을 통하여 기사의 성향을 분석한다. 해당 기사가 주식 시장에 반영되었을 경우 그 영향력을 분석하는 단계이다. 크게 긍정적, 부정적, 중립의 3가지 성향이 존재하게 된다. 오피니언 마이닝에서는 단어에 가중치를 두어 문장에 대한 점수를 계산하는 방법을 사용한다. 시스템 운용자가 단어 사전을 만들고 각 term들에 대한 가중치를 작성하게 된다. 각 term들에 부여된 가중치는 성향을 나타내게 되는데 부정의 부정, 긍정의 부정, 부정의 긍정 등 성향이 반대가 되는 경우를 위하여 긍정은 '+' 부호를 사용하고 부정은 '-' 부호를 사용하게 된다. 부호의 곱을 통하여 위와 같은 경우도 정확한 분류가 가능하게 된다.

문장에 대한 점수를 계산하기 위해서는 형태소 분석을 통해 문장 구조를 분석하고 각각의 관련된 문장 성분간의 가중치를 더하고 부호를 곱하는 방식으로 성향 분석을 진행한다. 기사에 대한 성향이 결정되면 기존의 주식 분석 시스템에 해당 기사에 대한 점수를 반영하게 된다.

4. 결론

본 논문은 유전 알고리즘과 기술적 분석을 사용하여 국내 주식 시장에서 매수와 매도를 추천하는 종목 추출 방법과 주식 시장 외적인 요소를 반영하는 방법에 대해 다루었다. 주식 시장에 영향을 미칠 수 있는 기사들을 문서 분류 단계에서 추출하여 신뢰성을 높이고, 매수 및 매도를 추천하는 종목에 기사의 성향을 분석하여 계산된 가중치를 적용한다. 따라서 기술적 분석 방법을 통해 구해진 신뢰도가 높은 데이터에 주식 시장 외적인 요소를 반영하여 더욱 안정적인 종목 추천을 제공하는 시스템으로서 긍정적인 가능성이 있음을 확인할 수 있었다. 뉴스 기사와 같은 매체는 주관적인 성격을 가지고 있기 때문에 기술적 분석으로 얻어지는 객관적인 데이터보다는 신뢰성이 낮은 편이다. 향후 기사 분석을 통해 나온 가중치를 주식 분석 시스템에 적용할 때의 적정 비율을 연구하여 시스템의 안정성을 증대시킬 것이다.

참고문헌

[1] Edwards, Robert D., Magee, John, Bassetti, "Technical Analysis of Stock Trends", American Management Association, pp.4-7, 1998.
[2] 양혜영, 빅데이터를 활용한 기술기획 방법론, 한국과학기술기획평가원, 제14호, 2012.
[3] Goldberg, D.E, Genetic Algorithm in Search, Optimization, and Machine Learning, Addison-Wesley, pp.10-17, 1989.
[4] 문병로, 쉽게 배우는 유전 알고리즘, 한빛미디어, pp.59-81, 2008.
[5] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol.34 ,No.1, pp.1-47, March 2002.
[6] 양정현, 명재석, 이상구, "상품 리뷰 요약에서의 문맥 정보를 이용한 의견 분류 방법", 정보과학회 논문지, 제36권, 제4호, pp.254-262, 2009.