섹터 내 동질성을 고려한 온라인 뉴스 기반 주가예측

저자
(Authors)

성노윤, 남기환

출처
(Source)

한국지능정보시스템학회 학술대회논문집 , 2017.11, 49-50(2 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07284540

APA Style

성노윤, 남기환 (2017). 섹터 내 동질성을 고려한 온라인 뉴스 기반 주가예측. 한국지능정보시스템학회 학술대회논문집, 49-50

섹터 내 동질성을 고려한 온라인 뉴스 기반 주가예측

성노윤
KAIST 경영대학
nyseong@business.kaist.ac.kr

남기환
한양대학교 경영대학
namkh@kaist.ac.kr

***Abstract** – 주가 예측이 활발히 진행되면서, 빅데이터를 결합한 주가 예측 연구도 활발히 진행되어지고 있다. 빅데이터 기반 언론의 효과를 접목시킨 연구 방법들은, 온라인 뉴스를 분석하여 주가 예측에 활용하는 연구가 주를 이루고 있다. 하지만 이러한 방법은 개인 회사에 대한 효과를 주로 살펴 보았고, 동질적인 섹터에 대한 효과를 살펴보는 연구는 있었지만, 이는 섹터 내에서도 이질성이 존재하는 등 실제 데이터 분석에 있어 여러 한계점을 가지는 것을 확인하였다. 본 연구는 이러한 기존 연구의 한계점을 데이터 마이닝 방법론을 적용하여, 주가에 영향을 미치는 기업의 동질적인 효과를 반영할 수 있는 방법론을 제안한다. 이를 기존에 연구되어 지고 있는 다양한 방법들과 비교 분석하여 본 연구의 우수성을 입증하였다.*

***Key Terms** – 주가예측, 딥러닝, 텍스트마이닝, 데이터마이닝, 컬러스터링.*

본 연구는 한국과학기술원의 미래선도형 특성화 연구사업 중, IoT 기반 초연결 사회를 위한 미래 비전에 관한 연구의 일환으로 이루어졌음

I. 서론

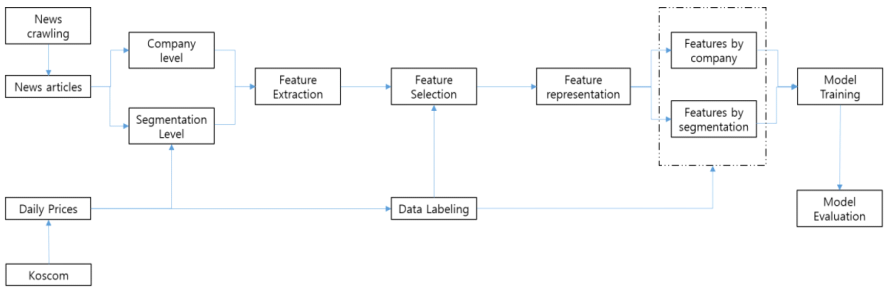
빅데이터 시대에 돌입하면서, 다양한 데이터를 기반으로 다양한 연구들이 진행되고 있다. 이러한 연구 흐름에 맞추어 기존 다양한 시도를 해왔던 주가예측 분야도, 빅데이터 분석을 활용한 연구가 활발히 진행되어 왔다. 가장 활발히 진행되고 있는 분야는 온라인 뉴스를 기반으로 각 기업의 주가를 예측하는 연구들 이었다.

하지만, 기존의 연구들을 보면, 대다수 각 개인 기업과 관련성이 가장 높은 기사들 만을 선택하여 연구를 한다. 관련성이 높은 기사들을 선택하는 방법은 일반적으로, 틱커(ticker)를 기반으로 선택하거나 (Shynkevich et al, 2016), 제목에 회사 이름이 포함되어 있거나(Peramunetilleke & Wong, 2002), 본문에 각 회사가 포함되어 있는 경우이다. 하지만, 카메라 제조 회사의 호재 뉴스가 나온다면 렌즈 부품 회사의 주가도 함께 오르듯이, 관련성이 높은 기업의 뉴스 또한 그 회사에 영향을 줄 수 있다. 이 사실을 이용하여, Shynkevich et al (2016)는 GICS 체계를 사용하여 관련성이 다른 기사들을 기반으로 예측을 하였고, 단순히 회사를 기반으로 하여 예측을 하는 것보다 더 정확하게 예측한다는 것을 보여주었다.

GICS는 산업체계를 반영하여 체계적으로 구성한 시스템이지만, 실제로 주가 간의 관련성을 반영한 체계라고 보기는 어렵다. 즉 기존 연구의 같은 섹터 내의 효과를 반영한 연구는 모든 섹터 내에는 모두 동질성을 가진다는 가정에서 출발하였다. 하지만 섹터 별로 보다 섹터 내에서 이질적인 성향을 띄는 섹터가 있고 보다 동질적인 성향을 띄는 섹터는 존재할 것이다. 하지만 기존 연구에서는 이러한 다양한 형태의 섹터의 특성을 모두 살펴보지 못하고 동질성을 띄는 섹터에 대해서만 연구를 진행하였다. 다양한 섹터를 확인해본 결과 기존 연구의 주장과는 상이한 결과를 나타내는 섹터도 존재 함을 알 수 있었다. 본 연구는 이러한 특성을 데이터 마이닝 기법을 접목시켜 섹터 내의 동질적인 패턴을 보다 명확하게 적용하기 위해 한번 더 분석을 수행할 수 있는 방법을 제안해 보고자 한다. 따라서, 본 논문에서는 GICS로 뉴스의 관련성을 측정하는 것을 대체할 방법으로 머신러닝 기법을 사용하여, 주식의 패턴에 따라 클러스터링을 하여, 관련성이 높은 기업들을 추출하여, 이를 기반으로 하여 주가를 예측하였다.

II. 데이터

본 논문에서 제시하는 방법을 실행하기 위해 실제 데이터를 가지고 실험을 하였다. 데이터는 2014년 1월 1일부터 2016년 12월 31일의 금융 뉴스와 주가 데이터로 이루어져있다. 뉴스는 한국 최대의 포털사이트인 네이버에 등록된 10개의 종합 신문과, 14개의 방송 통신과 9개의 경제 신문, 총 33개의 인터넷 뉴스의 모든 금융, 경제 관련 뉴스를 크롤링하였다. 이는 한국에서 대중이 접할 수 있는 대다수의 금융 뉴스를 포함한 정보로, 금융 뉴스가



<그림 1> 제한 모형

미치는 영향을 파악하기 좋은 데이터이다. 이 기간 동안 크롤링된 뉴스는 중복된 것을 제외하자 기사 수는 총 1,397,800 건이었었다. 뉴스 데이터의 형식은 카테고리(경제, 금융, 정치), 제목, 작성자, 포스트 시간, 내용 등이 있다.

III. 연구모형

본 연구에서는 섹터 내 이질성을 반영한 분석을 시행하기 위해 섹터 내에서 회사들 간에 세그멘테이션을 한다. 세그멘테이션을 시행하는 데는 다양한 방법이 있다. K-Means는 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로 각 클러스터와 거리 차이의 분산을 최소화 하는 방식으로 동작하며(MacQueen, 1967) DBSCAN은 클러스터들이 일정 이상의 밀도를 가지도록 하는 방법으로 클러스터링을 한다(Ester et al, 1996). 또한, CLARNAS(Ng and Han, 1994), BIRCH(Zhang et al., 1996) 등이 주로 사용된다. 그 중에서도 사용하기 간단하고, 많은 데이터를 처리하기 좋으며, 널리 사용되는 알고리즘은 k-menas이다. 따라서, 본 논문에서는 세그멘테이션을 K-Means Clustering을 사용하였다.

IV. 결론

정보의 양이 급증하게 되어 주식에 관련된 정보의 양이 무수히 많아지게 되는 빅데이터 시대에, 개인이 모든 뉴스를 읽고 주가에 영향을 미치는 정보만을 선별적으로 찾아내어 이용하는 것은 물리적으로 불가능해졌다. 이와 함께, 수 많은 텍스트 정보들을 자동적으로 처리하고 예측할 수 있는 알고리즘을 개발하는 것이 주요한 과제가 되었다. 특히, 수 많은 정보들 중에 영향을 주는 정보를 어떻게 선정하는 지도 주요한 과제가 되었다. 일반적으로는 제목에 회사의 이름이 있거나, 뉴스의 태그의 그 회사의 티커(ticker)가 있으면 그 회사에 영향을 주는 정보라고 인식을 한다. 본 연구에서는 영향의 범위를 각 회사와 동질적인 패턴을 보이는 그룹으로 확장을 하여 각 개인뿐만 아니라 영향력을 줄 수 있는 기업들도 함께 고려를 하여 예측을 할 때 어떻게 성능이 좋아지는 지에 대하여 연구를 하였다. 본 연구에서 제시한 방법을 Multiple Kernel Learning technique을 사용하여 예측을 한 결과 기존의 GICS 체계로 예측을 하거나, 개인 회사 단위로 예측을 하는 것보다 더 높은 예측률을 보였다.

V. 참고문헌

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., and Belatreche, A. Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. Decision Support Systems, Vol. 85(2016), 74-83.

Peramunetilleke, D., & Wong, R. K. Currency exchange rate forecasting from news headlines. Australian Computer Science Communications, Vol.24, No.2(2002), 131-139.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability Vol.1, No.14(1967), 281-297.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd Vol.96, No.34(1996), 226-231.

Ng, R. T., & Han, J. Efficient and effective clustering method for spatial data mining. In Proceedings of VLDB (1994), 144-155.

Zhang, T., Ramakrishnan, R., & Livny, M. BIRCH: an efficient data clustering method for very large databases. In ACM Sigmod Record Vol.25, No.2(1996), 103-114. ACM.