빅데이터 기반의 오피니언 마이닝을 이용한 기업 가치 평가 시스템 개발

이정태[O1], 천민아[1], 임상우[1], 전병석[1], 김재훈[2], 한영우[3]
한국해양대학교[1], 한국해양대학교[2], 한국예탁결제원[3]
{make8286[O], dkahffk0218, dlatkddn, dmdtnsky}@naver.com, jhoon@hhu.ac.kr, kim@mail.re.kr

Developing Corporate Valuation System with Opinion Mining Based on Big Data

Jung-Tae Lee[O1], Mina Cheon[1], Sang-Woo Lim[1],
Byung-Seok June[1], Jae-Hoon Kim[2], Yeong-Woo Han[3]
[1]Korea Maritime University, [2]Korea Maritime University, [3]Korea Securities Depository

요 약

빅데이터(Big Data)는 현재 생산되고 있는 데이터 중 그 규모가 방대하고, 생성 주기가 짧으며, 수치 데이터 뿐 아니라 텍스트 이외의 멀티미디어 등 비정형화된 데이터를 포함하는 대규모 데이터를 말한다. 빅데이터를 처리하여 가치 있는 정보를 추출하는 방법에 관한 연구가 활발하게 진행되고 있으며, 이를 바탕으로 빅데이터가 다양한 분야에서 활용되고 있다. 현재 국내 주식시장에서도 빅데이터를 이용하여 기업의 투자에 활용하고 있다. 이 논문에서는 인터넷의 증권과 관련된 뉴스를 수집하여 수집된 뉴스와 주가 지수를 이용하여 기업 뉴스 평가 시스템을 개발하는 방법을 제안한다.

주제어: 빅데이터, 오피니언 마이닝, 기업 뉴스 평가

1. 서론

빅데이터(big data)는 디지털 경제의 확산으로 규모를 가늠할 수 없을 정도로 많은 정보와 데이터가 생산됨에 따라 중요 키워드로 떠오르고 있다. 빅데이터란 정보 기술의 발전과 IT의 일상화가 진행되고 있는 현재 생산되고 있는 데이터 중 그 규모가 방대하고, 생성 주기가 짧으며, 형태도 수치 데이터뿐 아니라 텍스트 이외의 멀티미디어 등 비정형화된 데이터를 포함하는 대규모 데이터를 말한다[1]. 빅데이터가 등장함에 따라 데이터를 저장, 활용, 확산 및 공유하는데 그쳤던 과거와 달리 현재는 축적된 데이터 자원을 이용한 분석과 추론을 통해 해당 데이터로부터 가치창출을 하는 방향으로 데이터의 이용 흐름이 바뀌고 있다[2]. 이런 흐름 속에 빅데이터에 관한 분석기법까지 비약적으로 발전하면서 사회의 움직임을 더욱 정확하게 분석하고 예측할 수 있게 됨에 따라 빅데이터는 국토 보안, 의료, 마케팅, 증권업계 등 다양한 분야에서 활용되고 있다.

현재 국내 주식시장에서는 빅데이터로 뉴스를 선택하여 이를 분석해서 투자에 활용하고 있다. 모 증권사의 조사 결과에 따르면 일반적으로 투자자들은 한 기업에 대한 뉴스가 증가하면 관련 기업에 크게 관심을 갖게 되고 이는 주식 거래량과 주가 변동에 영향을 미치는 것으로 나타났다[3]. 이런 현상에 주목하여 특정 기업의 뉴스를 수집 후 오피니언 마이닝(opinion mining) 결과와 주가지수 정보를 기반으로 기업 가치를 평가하는 시스템을 개발하는 것을 목표로 한다.

이 시스템은 특정 기업의 뉴스를 수집하는 웹 크롤링 에이전트와 수집된 기업뉴스에 대한 평가를 수행하기 위한 오피니언 마이닝 엔진, 기업 평가 시 활용할 수 있는 가중치를 조정할 수 있는 기능과 평가된 결과를 조회 할 수 있는 화면, 오피니언 마이닝 결과와 주가지수 정보를 기반으로 기업 가치를 평가하는 예측 엔진으로 구성된다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 이 시스템을 개발하는데 필요한 관련 연구를 살펴보고, 3장에서는 제안하는 기업 평가 시스템에 대해 설명하고, 마지막으로 4장에서 결론 및 향후 연구를 기술한다.

2. 관련 연구

2.1 오피니언 마이닝

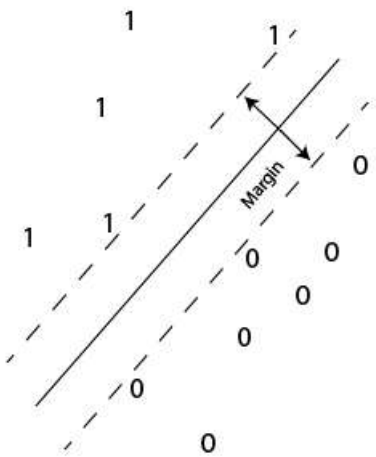
빅데이터를 분석하는 기술과 방법들은 기존 통계학과 전산학에 사용되던 데이터 마이닝(data mining), 기계학습, 자연 언어 처리, 패턴 인식 등이 있다[4]. 오피니언 마이닝은 텍스트 마이닝(text mining)의 한 분야로 소셜 미디어(social media) 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별하는 기술이다[5]. 일반적으로 오피니언 마이닝은 특정 주제에 대한 사람들의 주관적인 의견들을 모아 문장 단위로 분석한다. 문장 분석에서는 사실과 의견 부분을 구분한 후, 의견 부분을 토대로 긍정과 부정으로 나눈 뒤 그 강도를 측정한다.

오피니언 마이닝은 대체로 크게 특징 추출(feature extraction), 의견 분류(opinion classification), 요약 및 표현 등의 가지 세 단계로 이루어진다. 첫째, 특징 추출 단계는 오피니언 마이닝을 수행함에 있어 유용한 정보라고 판단되는 여러 특징(feature)들을 정의하고 추

출해 내는 단계이다. 이때 단순히 특징만을 추출하는 것이 아니라 주제어에 대하여 긍정적인인지 부정적인지에 대한 극성 판별도 함께 이루어진다. 둘째, 의견 분류 단계에서는 대상의 특징을 중심으로 문서를 요약하고 추출된 특징과 의견을 나타내는 어휘가 주제어에서 어떤 의미로 사용되었는가에 대한 판단 및 분류를 하는 단계이다. 셋째, 요약 및 표현단계에서는 주제어에 대한 사용자의 의견들 중 의견 성향이 밝혀진 정보들을 요약하여 전체 정보의 내용을 효율적으로 사용자에게 보여주는 단계이다[6-9].

2.2 SVM

SVM(Support Vector Machine)은 감독 기계 학습 기법(supervised learning)의 한 종류로써 데이터를 분류하는데 좋은 성능을 보인다. SVM이 좋은 서능을 보이는 가장 큰 이유는 일반적으로 학습 시에는 매우 많은 특징을 다루어야 하는데, SVM의 경우 특징의 수에 의존하지 않는 over-fitting 방지 알고리즘을 가지고 있어 많은 수의 특징 공간을 다룰 수 있다. SVM의 기본 원리는 많은 통계 학습 모델 중에서 가장 높은 성능 결과 값을 가지고 있는 모델로, 두 개의 클래스의 구성 데이터들 가장 잘 분리할 수 있는 결정함수(hyperplane)를 찾는 것이다[10-12].



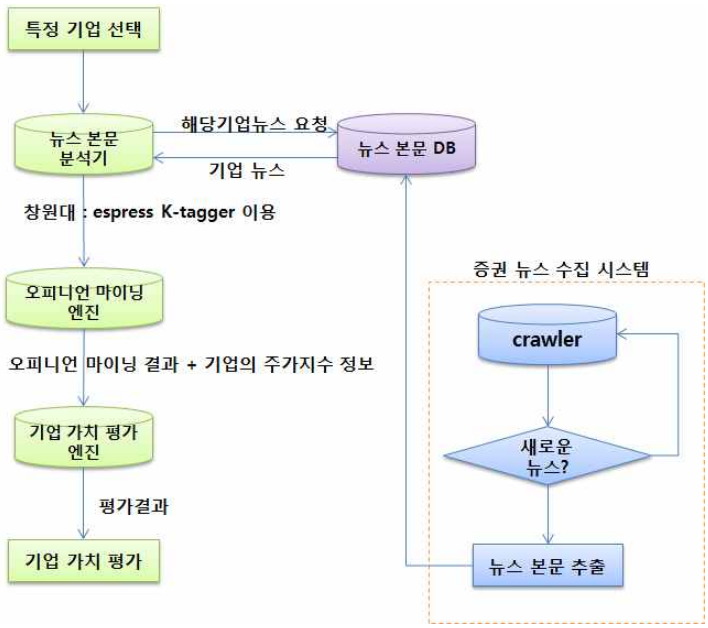
(그림 1) 선형 SVM의 결정함수[13]

이 논문에서는 긍정/부정의 2가지 클래스를 판별하고자 하며 Weka의 SVM[15]을 사용한다. 실험 데이터의 추출은 증권 뉴스 기사들을 모아서 사람이 직접 분류하여 데이터를 수집한다. 성능 평가를 위한 자질 단어의 수집도 함께 이루어졌다. 내용어의 수집은 뉴스 기사 상에서 형태소 분석을 통해 의미 있는 단어들을 선별하였으며, 문서에서 지대한 영향을 끼칠 수 있는 감정단어 선별은 국립국어원에서 제공하는 세종 말뭉치의 형용사를 기반으로 감정단어 사전을 구축하였다.

3. 기업평가 시스템

3.1 전체 시스템

이 논문에서 제안하는 전체 시스템의 동작 구조는 (그림 2)와 같다. 먼저 특정 기업의 뉴스를 수집하는 웹 크롤링 시스템을 통해 증권 사이트에서 관련 뉴스들을 수집한다. 새로운 뉴스를 수집했을 경우 html 태그를 제거하여 뉴스 본문을 추출한 후, 추출한 본문 데이터베이스에 저장한다. 사용자가 특정 기업을 선택하게 되면, 뉴스 본문이 저장되어 있는 데이터베이스에 접속하여 해당 기업의 뉴스들을 창원대에서 제공하는 형태소분석기인 Espresso-K-Tagger[16]를 이용하여 형태소 분석을 수행한다. 수행된 결과를 바탕으로 수집된 기업뉴스에 대한 평가를 위해 오피니언 마이닝 엔진을 이용하여 오피니언 마이닝을 수행하고, 그 결과와 주가지수 정보를 기반으로 기업 가치를 평가하는 예측 엔진을 사용하여 그 결과값을 화면에 조회할 수 있도록 한다.



(그림 2) 전체 시스템

3.2 오피니언 마이닝 엔진

수집한 뉴스의 본문을 형태소 분석하여 이를 긍정/부정으로 분류한다. 이를 위해 문서를 긍정/부정으로 나뉠 때 일반적으로 사용하는 기계학습 알고리즘 중 SVM을 이용한다. 감정단어는 세종말뭉치를 기반으로 '좋다', '나쁘다'는 감정단어의 유의어를 기반으로 확장시켜 수집하고 내용어는 형태소 분석을 통해 분석한 내용 중 불용어를 제외한 전체 문서의 분석결과를 토대로 감정단어를 제외한 명사, 형용사, 부사, 동사의 단어들을 수집한다. 이렇게 수집한 감정단어, 내용어, 감정단어+내용어를 토대로 SVM을 이용하여 문서를 분류한다. 오피니언 마이닝에서 감정단어와 관련된 자질들의 가중치가 중요하므로, 감정단어의 경우 $TF-IDF$를 이용하여 다른 자질보다 가중치를 더 높게 부여한다.

$TF-IDF$는 문서에 단어 t가 나타난 단어의 빈도수 tf_t와 문서의 역문서 빈도수인 idf_t를 이용하여 가중치를 구하는 방법이다. idf_t를 구하는 식은 (1)과 같다.

$$idf_t = \log_2 \frac{N}{df_t} \quad (1)$$

여기서 N은 전체 문서의 수이며, df_t는 단어 t가 출현한 문서의 빈도수이다.

$TF-IDF$는 여러 문서로 이루어진 문서군이 존재할 때 단어 t가 특정 문서에서 얼마나 중요한 것인지를 나타내는 통계적 수치다. $TF-IDF$를 구하는 방법은 식 (2)와 같이 tf_t와 idf_t를 곱하는 것으로 구할 수 있다.

$$TF-IDF = tf_t \times idf_t \quad (2)$$

각 감정 단어 자질 t마다 출현 횟수를 이용하여 식 (3)을 이용하여 가중치 W_t를 부여한다. 이는 각각의 긍정과 부정의 의미에 대하여 많이 출현한 단어의 의미가 더 강하므로 높은 가중치를 부여하고, 전체에 비례하여 평준화시키는 것이다.

$$W = \frac{t_{cnt}}{N_{\max cnt}} \quad (3)$$

t는 감정 단어 자질이며 t_{cnt}는 해당 단어가 뉴스에 출현한 횟수이며 $N_{\max cnt}$는 감정 자질의 단어들 중 최대로 출현한 횟수다. 각 감정 자질의 가중치는 식 (2)와 (3)을 이용하여 구한다[14].

3.3 기업 가치 평가 엔진

오피니언 마이닝의 결과와 수집한 주가지수 정보를 바탕으로 기업 가치를 평가한다. 해당 기업과 관련된 수집된 뉴스의 오피니언 마이닝 수행결과 긍정 뉴스가 부정 뉴스보다 더 많을 경우 기업의 가치를 높게 평가하는 것을 기본으로 하며, 여기에 수집한 주가지수도 평가 항목으로 적용한다. 대체로 우량기업의 경우 뉴스의 평가가 기업 가치에 미미한 영향을 미치는 경우가 많으므로 [15], 이 경우에는 해당 기업이 가지고 있는 주가지수에 더 많은 가중치를 줘서 평가하도록 한다.

4. 결론 및 향후 연구

오피니언 마이닝은 기본적으로 영화나 특정 제품의 상품평 등 사람들의 주관이 들어간 내용에 더 적합하기 때문에, 객관적인 사실이 대부분의 내용인 뉴스의 오피니언 마이닝에는 한계점이 존재한다. 따라서 차후 뉴스를 수집할 때 해당 뉴스의 댓글들을 따로 수집하여 이를 평가 항목에 추가하는 방향으로 연구를 진행하면 지금 구성하고 있는 시스템보다 더 좋은 결과를 낼 수 있을 것으로 예상된다.

참고 문헌

[1] 정용찬, "빅데이터", 커뮤니케이션북스, pp.2, 2012.

[2] 한국정보화진흥원, 빅데이터 전략 연구센터, "새로운 미래를 여는 빅데이터 시대", pp.19, 2013.

[3] 김수진, "[증시閑담] 뉴스 빅데이터, 주식을 분석해라", Chosun Biz, http://biz.chosun.com/site/data/html_dir/2012/12/05/2012120500308.html, 2012.

[4] 정병권 외 2명, "미래사회와 빅 데이터(Big data) 기술", IT기획 시리즈, 정보통신산업진흥원, pp. 20-21, 2012.

[5] 최종후, "빅 데이터 시대가 도래한다", Korea University Sejong Campus Magazine, vol.20, pp.9, 2012.

[6] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena, "Large-Scale Sentiment Analysis for News and Blogs," Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2007), 2007.

[7] E. Boiy, P. Hens, K. Deschacht, and M. Moens, "Automatic Sentiment Analysis in On-line Text," Proceedings of the ELPUB2007 Conference on Electronic Publishing, June 2007.

[8] J. Yi and W. Niblack, "Sentiment Mining in Web-Fountain," Proceedigns of the International Conference on Data Engineering (ICDE'05), pp. 1073-1083, 2005.

[9] T. Nasukawa, J. Yi, "Sentiment analysis: capturing favorability using natural language processing," Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture, pp. 70-77, 2003.

[10] Y. Bao, and N. Ishii, "Combining Multiple K-Nearest Neighbor Classifiers for Text Classification by Reducts." Proceeding of the fifth International Conference on Discovery Science, pp. 340-347, 2002.

[11] 이재식, 이종운, "사례기반 추론을 이용한 한글 문서분류 시스템", 경영정보학연구, 제12권, 제2호, 2002.

[12] 김진상, 신양규, "베이지안 학습을 이용한 문서의 자동분류", Journal of the Korean, Data & Information Science Society, Vol. 11, No. 1, pp 19-30, 2000.

[13] Chris Thornton, "Machine Learning - Lecture 15 Support Vector Machines", http://www.sussex.ac.uk/Users/christ/crs/ml/lec08a.html, 2011.

[14] 황재원, 고영중, "감정 분류를 위한 한국어 감정 자질 추출 기법과 감정 자질의 유용성 평가", 인지과학, 제19권, Vol.4, pp. 506-507, 2008.

[15] Weka, http://www.cs.waikato.ac.nz/ml/weka/

[16] 창원대학교 AIR 연구실, Esspreso K Tager, http://air.changwon.ac.kr/blog/2012/01/04/esspreso-pos-tagger-for-korean/