온라인 주식 포럼의 핫토픽 탐지를 위한 감성분석 모형의 개발

Development of Sentiment Analysis Model for the hot topic detection of online stock forums

저자
(Authors)

홍태호, 이태원, 리징징
Taeho Hong, Taewon Lee, Jingjing Li

출처
(Source)

지능정보연구 22(1), 2016.03, 187-204(18 pages)
Journal of Intelligence and Information Systems 22(1), 2016.03, 187-204(18 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06646758

APA Style

홍태호, 이태원, 리징징 (2016). 온라인 주식 포럼의 핫토픽 탐지를 위한 감성분석 모형의 개발. 지능정보연구, 22(1), 187-204

ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

J Intell Inform Syst 2016 March: 22(1): 187~204 http://dx.doi.org/10.13088/jiis.2016.22.1.187

온라인 주식 포럼의 핫토픽 탐지를 위한 감성분석 모형의 개발*

홍태호
부산대학교 경영학과
(hongth@pusan.ac.kr)

이태원
부산대학교 중국연구소
(twanny@pusan.ac.kr)

리징징
부산대학교 중국연구소
(jingjing@pusan.ac.kr)

소셜 미디어를 이용하는 사용자들이 직접 작성한 의견 혹은 리뷰를 이용하여 상호간의 교류 및 정보를 공유하게 되었다. 이를 통해 고객리뷰를 이용하는 오피니언마이닝, 웹마이닝 및 감성분석 등 다양한 연구분야에서의 연구가 진행되기 시작하였다. 특히, 감성분석은 어떠한 토픽(주제)를 기준으로 직접적으로 글을 작성한 사람들의 태도, 입장 및 감성을 알아내는데 목적을 두고 있다. 고객의 의견을 내포하고 있는 정보 혹은 데이터는 감성분석을 위한 핵심 데이터가 되기 때문에 토픽을 통한 고객들의 의견을 분석하는데 효율적이며, 기업에서는 소비자들의 니즈에 맞는 마케팅 혹은 투자자들의 시장동향에 따른 많은 투자가 이루어지고 있다. 본 연구에서는 중국의 온라인 시나 주식 포럼에서 사용자들이 직접 작성한 포스팅(글)을 이용하여 기존에 제시된 토픽들로부터 핫토픽을 선정하고 탐지하고자 한다. 기존에 사용된 감성 사전을 활용하여 토픽들에 대한 감성값과 극성을 분류하고, 군집분석을 통해 핫토픽을 선정하였다. 핫토픽을 선정하기 위해 k-means 알고리즘을 이용하였으며, 추가로 인공지능기법인 SOM을 적용하여 핫토픽 선정하는 절차를 제시하였다. 또한, 로짓, 의사결정나무, SVM 등의 데이터마이닝 기법을 이용하여 핫토픽 사전 탐지를 하는 감성분석을 위한 모형을 개발하여 관심지수를 통해 선정된 핫토픽과 탐지된 핫토픽을 비교하였다. 본 연구를 통해 핫토픽에 대한 정보 제공함으로써 최신 동향에 대한 흐름을 알 수 있게 되고, 주식 포럼에 대한 핫토픽은 주식 시장에서의 투자자들에게 유용한 정보를 제공하게 될 뿐만 아니라 소비자들의 니즈를 충족시킬 수 있을 것이라 기대된다.

주제어 : 감성분석, 오피니언 마이닝, SM, 핫토픽, 온라인 포럼

논문접수일 : 2015년 9월 9일 논문수정일 : 2016년 3월 15일 게재확정일 : 2016년 3월 16일
원고유형 : 일반논문 교신저자 : 이태원

1. 서론

Web 2.0 시대에 들어와서 인터넷의 사용증가로 인해 사용자가 직접 작성한 글들이 폭발적으로 늘어나기 시작하였다. 소셜 미디어를 통해 사용자들은 자신이 직접 작성한 의견 혹은 리뷰를 남기고, 이를 통해 다른 사용자와 상호간의 교류 및 정보를 공유하게 되었다. 사용자들은 일반적으로 블로그, 위키, 마이크로블로그, 온라인 포럼 등 다양한 방식으로의 접근으로 방대한 정보가

* 이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A5A2A01015166)

생성되고 있으며, 이에 대해 비구조적인 텍스트 기반의 분석이 요구되고 있는 웹마이닝, 오피니언마이닝 혹은 감성분석이 주목을 받고 있다 (Chen et al., 2012; Pang et al., 2002; Wang et al., 2014).

온라인상의 텍스트를 분석하는 방법들 중의 하나인 감성분석은 텍스트 문서에 내포되어 있는 정보들을 이용하여 주관적인 정보를 추출하는 기법이다. 감성분석은 어떤 특정한 토픽(Topic)에 대해 말하는 사람들과 글을 작성한 사람들의 태도, 입장 및 감성을 알아내는데 목적을 두고 있다(Tan et al., 2008). 사용자들은 어떠한 주제를 통해 직접 작성한 포스팅(글)을 작성함으로써 기업에서는 이를 통해 소비자들의 니즈에 맞는 마케팅 혹은 주식시장에서의 투자자들에게 최신 동향에 대한 흐름을 알 수 있도록 소비자들에게 유용한 정보를 제공한다(Zhang et al., 2009). 특히, 온라인 포럼은 다양한 사람들과 전문가들이 참여하여 토론 방식의 특정 토픽에 대한 포스팅을 작성하고, 자신의 생각과 의견에 대한 정보를 제공함으로써 상호간의 교류를 활발히 할 수 있는 곳이다. 주식시장의 경우 투자자들은 온라인 포럼을 이용하여 많은 유용한 정보들을 이용하여 사용자들과의 교류를 통해 최신 정보 및 동향을 파악하여 수익을 창출시킬 수 있는 계기가 된다. 온라인 주식 포럼은 사용자들이 당일 주식 시장에서 발생한 주가 변동이나 시장 경제에 관한 정부 정책 발표, 토픽에 따른 자신의 경험이나 취향 등 다양한 내용으로 정보를 제공하고 있다. 전문가들은 사용자들이 직접 작성한 포스팅을 분석하여 주식시장의 예측 및 분석에 대한 글을 작성하여 공유함으로써 사용자들과의 유대관계 형성뿐만 아니라 다양한 정보를 제공한다. 특히, 중국 주식시장의 강한 성장세로 인해 투자자들이 증가하여 많은 사용자들이 관심을 가지기 시작하여 많은 정보들이 생성되고 있다. 중국의 온라인 시나 주식 포럼에서는 다양한 토픽으로 구성되어 있어 많은 사용자들이 이용을 하고 있으며, 가장 많은 관심을 보이는 토픽인 핫토픽(Hot Topic)을 탐지한다면 사용자의 관심을 유발시키는 토픽이 무엇인지 파악 할 수 있게 된다.

본 연구에서는 온라인 주식 포럼에서 토픽을 통해 생성되는 사용자들의 정보들을 이용하고, 온라인 주식 포럼에서 기존에 제시된 다양한 토픽들 중 핫토픽을 탐지하는데 초점을 두어 감성분석에 대한 연구를 진행하고자 한다. 먼저, 온라인 주식 포럼에서 데이터를 추출하여 기존에 제시된 토픽들을 이용하여 핫토픽을 선정한 후 기계학습기법을 통해 핫토픽이 사전 탐지가 가능한 지를 파악한다. 핫토픽의 선정과 핫토픽의 사전 탐지에 대한 결과를 비교하기 위해 관심지수를 이용하여 비교 및 분석을 하고자 한다. 또한 어휘기반 접근방법인 감성사전의 사용과 기계학습 접근방법을 이용하여 통합 접근방법의 연구를 진행하고자 한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 선행연구를 제시하고, 제 3장에서는 핫토픽 탐지를 위한 연구 프레임워크를 제시하며, 제 4장에서는 모형에 대한 실험결과를 제시한다. 마지막으로 제 5장에서는 결론 및 향후 연구방향에 대해 제시한다.

2. 선행 연구

2.1 감성분석

감성분석(Sentiment Analysis)은 텍스트로부터 어떻게 감성이 표현되고 있는지에 대한 분석과 특정한 주제에 대해 긍정적 혹은 부정적인 표현

들이 포함되어 있는지를 분석하는 방법이다(Hu and Liu, 2004; Hu et al., 2012; Jin et al., 2014). 감성분석은 오피니언마이닝(Opinion mining)이라고도 불리며, 텍스트마이닝의 한 부분으로써, 기술적인 방법을 통해 텍스트 데이터로부터 포함되어 있는 오피니언을 추출하고 가치 있는 정보를 추출해 내는 것으로 정의한다(Liu, 2012). 감성분석 어휘를 기반으로 접근하는 방법(Lexicon-based Approach), 기계학습을 기반으로 접근하는 방법(Machine Learning Approach), 어휘기반 접근방법과 기계학습 접근방법을 통합하여 접근방법(Hybrid Approach) 등 3가지로 나누어 분석이 이루어진다(Medhat et al., 2014).

어휘기반 접근방법은 긍정적 단어와 부정적 단어에 대해 어떤 의미가 있는지 분석하는 방법으로 미리 편집된 감성단어와 잘 알려진 감성어휘를 사용하는 것이 대표적이다((Turney and Littman, 2003; Martin-Valdivia et al., 2013; An and Kim, 2015). 어휘기반 접근방법 중 사전기반 방법(Dictionary-based Approach)은 수집된 텍스트를 분석하여 문서내의 단어들을 추출하고 사전을 구축하여 새로운 문서가 주어졌을 때 기존에 만들어진 사전과 비교하여 문서를 분류하는 방식이다(Maks and. Vossen, 2012; Oh and Kang, 2013).

기계학습 접근방법은 수집된 데이터를 기반으로 학습시킨 후 모형을 통해 분석을 하여 미래를 예측하는 기법을 말하며 사용되는 알고리즘과 언어적 변수를 사용하여 문서 분류문제로 감성분석을 해결하기 위해 알고리즘에 의존하는 성향이 있다(Medhat et al., 2014). 기계학습 접근방법은 지도학습 기법(Supervised Learning method)과 비지도학습기법(Unsupervised Learning method)으로 분류되며 다양한 데이터마이닝 기법들을 활용할 수 있다.

2.1.1 감성사전을 이용한 어휘기반 접근방법

감성분석은 단어의 극성에 대한 특징에 따라 분류되며, 일반적으로 기존에 구축된 사전을 통해 단어에 점수를 부여하여 긍정적인지 부정적인지를 분류할 수 있다(Pang and Lee, 2008). 감성분석을 위해 사용되는 감성사전의 경우 영어 감성사전으로 SentiWordNet, 중국 감성사전으로 HowNet의 감성사전, 간체 중국어 극성 사전 등을 이용하여 다양한 연구들이 진행되고 있다(Baccianella et al., 2010; Yu et al., 2013; Li and Wu, 2010). 한국의 경우 공신력 있는 한국어 감성사전이 제공되지 않고 있어 각 연구별로 감성사전을 구축한 후 감성분석을 진행하고 있다(Kim et al., 2014).

중국 국가지식기반시설(National Knowledge Infrastructure) HowNet이 발표한 '감성분석 용어사전'은 긍정, 부정의 양극성 단어를 구분해 줄 뿐만 아니라 단어의 강도를 표현하는 사전 또한 제공되고 있어 편리하게 사용되고 있다. 현재 중국에서는 HowNet의 감성사전을 기반으로 이용하는 연구가 점차 늘어나고 있다.

Li and Wu(2010)은 HowNet의 감성사전과 단어의 감도를 표현하는 사전 모두를 이용하여 수집한 문서에 점수를 부여하고 계산을 통해 온라인 시나 스포츠 포럼의 핫토픽을 탐지하는 연구를 진행하였다.

Zhang et al.(2009)은 단어 의존성을 기반으로 문장의 감성을 결정하고 문장을 취합하여 문서의 감성을 예측하기 위해 중국문장과 중국기사에 대해 감성의 극성을 예측하기 위해 단어 의존성 구조를 고려한 규칙기반의 의미론적 분석방법을 제안하였다. 전자상거래 웹사이트에서 공공 보건 문제와 제품 리뷰 등 다양한 토픽 영역에서 중국 기사의 감성에 대한 극성을 예측하는데 초점을 두었다. SVM, 나이브 베이즈, 의사결

정나무기법 중 C4.5와 규칙기반 접근방법을 이용하여 실험을 통해 예측 성과를 비교하는 연구를 진행하였다.

2.1.2 데이터마이닝을 이용한 기계학습 접근방법

감성분석에 사용되는 데이터마이닝 기법은 지도학습기법과 비지도학습기법으로 나누어서 설명할 수 있다. 지도학습기법에서 주로 사용되는 로짓, 의사결정나무, SVM 기법들에 대해 설명하면 다음과 같다.

로짓분석은 어떤 사건에 대한 발생 유무를 직접 예측하는 것이 아니라, 그 사건이 발생할 확률을 예측하고, 0과 1 사이의 결과값을 가지게 된다. 로짓분석 결과 종속 변수의 값이 분류기준값 이상이면 사건이 일어나고, 이하이면 사건이 일어나지 않는 것으로 예측하게 된다.

의사결정나무는 나무와 같은 구조를 사용하며 분류, 군집, 변수선택, 예측문제 등을 위해 사용한다. 또한, 예측을 수행하는 과정이 나무구조에 의한 추론규칙(induction rule)으로 표현되어 과정 및 결과를 해석하는데 용이하다는 장점을 가지고 있다. 의사결정나무를 만드는 데에는 CART(Breiman et al., 1984), CHAID(Kass, 1980; Hartigan, 1975), C4.5(Quinlan, 1986)과 같은 다양한 알고리즘들이 있다.

CART(Classification and Regression Tree) 알고리즘은 가장 많이 사용되는 방법 중 하나로 각 마디에 하나의 독립변수의 값에 따라 데이터를 두 개로 나누는 나무구조를 만들며, 이분형으로 표현할 수 있는 특성에 기반하여 분할을 시행한다.

CHAID(Chi-squared Automatic Interaction Detection)는 Hartigan(1975)의 제안으로 카이제곱 통계량(이산형 목표변수) 혹은 F-검정(연속형 목표변수)를 이용하여 다지분리(Multiway Split)를 수행하는 알고리즘이다. 다지분리란 부모마디에서 자식마디들이 생성될 때 2개 이상의 분리가 일어나는 것을 의미하며, CHAID는 목표변수에 가장 유의한 독립변수를 찾아 변수기준으로 나무를 형성한다(Park and Cho, 2004). C4.5는 Quinlan(1993)에 의해 ID3(Interative Dichotomizer 3) 알고리즘의 단점을 보완하고 새로운 기능을 추가한 알고리즘으로써, 각 마디에서 여러 개의 분리구조로 의사결정나무를 만든다.

SVM(Support Vector Machine)은 Vapnik(1995)이 제안한 학습이론으로 기존의 통계적 이론에서 이용되는 경험적 위험 최소화 원칙(Empirical risk minimization)을 이용하여 일반화 오류를 줄여 패턴 인식과 문서 범주화 등에서 우수한 성능을 보여주고 있다(Burges, 1998). SVM은 서로 다른 값을 가지는 두 개의 클래스를 포함한 학습집합을 최대한으로 분류하는 초평면을 설정하여 분류하며, 이진분류에 있어서 다른 분류기법들과 비교하였을 때 우수한 성능을 보인다고 알려져 있다.

비지도학습기법에서 주로 사용되는 군집분석 중 k-means와 SOM에 대해 살펴보기로 한다.

군집분석(clustering analysis)은 데이터에 대해 주어진 관측값을 사용하여 전체를 몇 개의 유사집단으로 분류하는 기법으로 k-means와 SOM 알고리즘이 가장 많이 사용되고 있다.

k-means 알고리즘은 간단한 구조를 가지고 있어 다양한 형태의 데이터에 적용이 가능하고, 각 군집의 중심에서 개체까지의 거리 차이 분산을 최소화하는 방식으로 계산된다. k-means는 개체들을 k개의 초기 군집으로 시작하여 군집의 수와 최초 시작점을 지정한 후 각 시작점으로부터 모든 개체와의 거리를 계산하여 가장 거리가 가까운 개체를 중심으로 초기 군집을 이룬다. 각

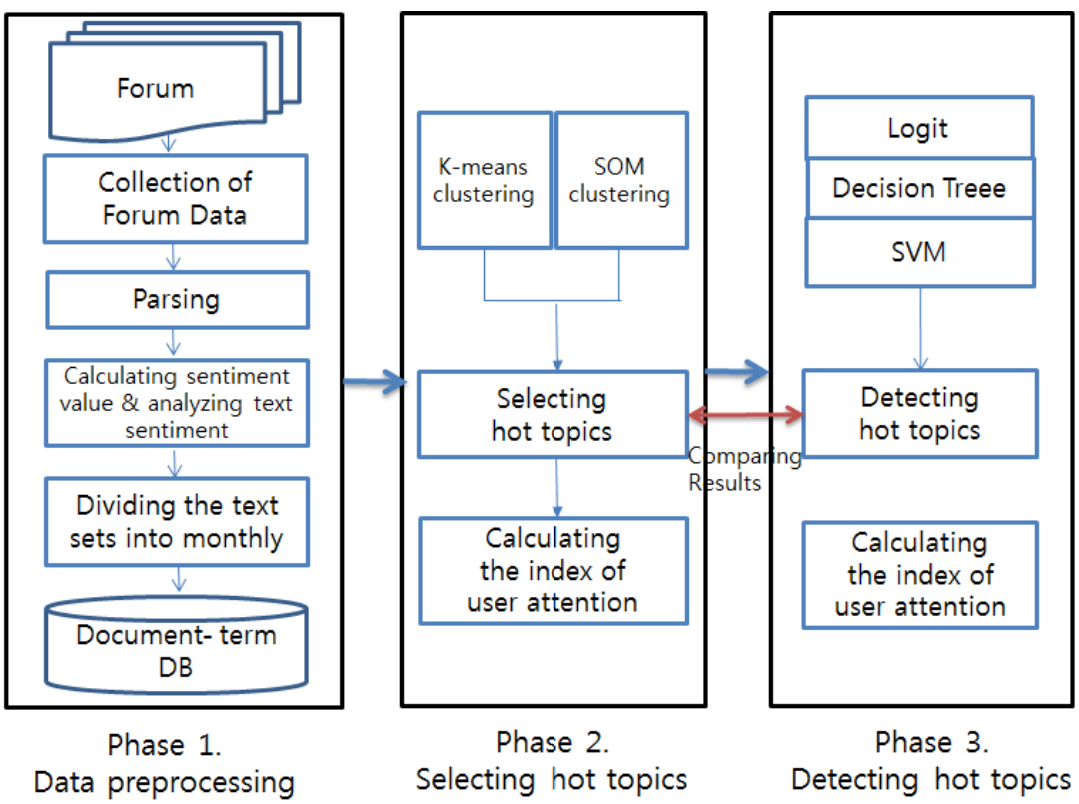
군집 내에서의 중심점을 찾은 후 모든 개체들과의 거리를 계산하고 가장 근거리에 있는 개체를 중심으로 새로운 군집을 생성하게 되며, 더 이상의 군집에 대한 형태가 변하지 않을 때까지 반복하여 진행된다(Hong and Kim, 2010).

SOM(self-organizing map)은 Kohonen(1988)에 의해 제안된 비지도학습기법의 인공신경망 모델 중의 하나로 군집 특성을 이용한 기법이다. SOM 알고리즘은 주어진 데이터들로부터 규칙을 찾아내는데 목적을 두고 있으며 정확한 해답을 제공하지 않고 자기 스스로 학습할 수 있는 능력을 가지고 있다. 또한, 구조적으로 수행이 빠른 알고리즘으로 입력 데이터들의 분포가 시간에 따라 스스로 조직화를 통한 정확한 군집 기법으로 알려져 있다.

2.2 주식 분석을 위한 감성분석 연구

주식에 대한 감성분석 연구가 많이 진행되어 왔으며, 일반뉴스, 금융뉴스, 회사뉴스, 회사년도 보고서, 트위터 등 미디어에서 나온 데이터를 수집하여 분석함으로써 주가예측이나 주식시장 동향 등 다양한 데이터를 통해 감성분석을 시도하였다.

Schumaker et al.(2012)은 금융뉴스제목에 대한 감성을 평가하여, 기사의 어조(주관적/객관적)가 미래 주가 추세에 어떠한 영향을 끼칠 수 있는지와 주관적인 의견(긍정/부정)이 미래 주가 추세에 어떠한 영향을 미칠 수 있는지를 조사하였다.



〈Figure 1〉 A Framework for Hot Topic Recommendation

Jin et al. (2013)는 일반뉴스 내용과 과거 통화 시장지수를 이용하여 통화시장의 지수변동을 예측하였다. 또한 Fung et al.(2003)는 회사뉴스를 이용하여 다중 시계열 방법으로 주가예측에 대한 연구를 하였고, Huang et al.(2010)는 대만 주식시장에 대한 실시간 뉴스를 분석함으로써 주식매매를 통해 의사결정에 도움이 될 수 있다는 것을 증명하였다. Bollen and Huina(2011)는 트위터에서 작성된 댓글을 분석함으로써 사람들의 감정을 파악하여 주식시장에 반영되는 것으로 주가변동 추이를 예측하였다.

3. 연구모형

3.1 연구 프레임워크

본 연구에서는 온라인 주식 포럼에서 추출한 데이터를 사용하여 기존에 제시된 토픽들로부터 핫토픽을 선정하고 탐지하여 최종적으로 선정된 핫토픽(Hot Topic)을 추천하고자 <Figure 1>과 같이 개발하였다.

3.1.1 데이터 전처리 단계

온라인 주식 포럼에서 토픽에 대한 포스팅을 추출하기 위해 텍스트 수집 프로그램을 이용하여 데이터를 수집한다. 수집된 데이터는 작성자, 제목, 작성시간, 조회 수, 내용 등으로 구분되어 있으며 토픽별로 문서들을 분류한 후 문서 내에 포함되어 있는 문장들을 단어 단위로 프로그램 'R'을 이용하여 분할시킨다. 본 연구에 사용될 데이터는 중국어로 되어 있으며, 영어나 한국어와 다르게 한 문장 안에 띄어쓰기가 전혀 되어 있지 않아 전처리과정을 통해 문장을 단어별로 분할하여 저장하여야만 한다. 감성분석을 위해 분할된 단어들에 대한 감성값을 계산하기 위해 HowNet의 감성 용어 사전을 이용하였으며, 용어 사전에 수록된 어휘의 극성을 5가지로 분류하여 감성값을 계산한다. 각 어휘의 극성을 구분하기 위해 1점, 3점, 5점, 7점, 9점으로 설정하고, 문서 내에 포함되어 있는 모든 단어들의 점수에 대한 합을 구하여 문서의 감성값으로 간주한다. 또한 문서의 극성을 분류하기 위해 감성사전을 이용하여 긍정 단어이면 +1점, 부정 단어이면 -1점을 부여하고 양의 점수를 가진 문서는 긍정문서, 부의 점수를 가진 문서는 부정문서, 0점이면 중립에 해당하는 문서로 분류한다. 최종적으로 분류된 문서들은 월별로 문서를 분류한 후 문서-용어 데이터베이스에 저장한다.

3.1.2 핫토픽 선정 단계

핫토픽이란 온라인 주식 포럼을 이용하는 사용자들로부터 기존에 사용된 토픽들 중 가장 많은 관심을 유도한 토픽을 의미한다. 핫토픽을 선정하기 위해 먼저 월별로 분류된 문서들을 그룹화하고, 토픽을 군집화하여 k-means와 SOM 알고리즘을 통해 핫토픽을 탐지한다. 군집 결과에 대한 정확성을 향상시키기 위해 두 가지 기법에 사용되는 입력변수는 총 5개로 <Table 1>과 같이 나타낼 수 있다. 핫토픽의 선정 기준은 군집 시 각 클러스터에서 중심과의 거리가 가장 가까

〈Table 1〉 Input Variable for Hot Topic Detection

Variable	Description
NUM_{ij}	No. of topic post
$CLICKS_{ij}$	Average No. of clicks of topic post
SEN_{ij}	Average sentiment value of topic post
POS_{ij}	No. of positive post
NEG_{ij}	No. of negative post

i: Month, j: Topic

운 토픽이 클러스터를 대표할 수 있는 핫토픽으로 선정한다.

핫토픽을 선정하기 위해 5개의 입력변수를 이용하여 월별 문서 집합을 군집화하고, 각 군집 내의 중심점을 찾은 후 모든 토픽들과의 거리를 계산한다. 중심점에 있는 토픽과 중심점과의 거리가 가장 가까운 토픽을 핫토픽으로 선정하고, 월별로 선정된 핫토픽 수를 전체 기간에 선정된 핫토픽 수의 합인 관심지수로 산출한다. 본 연구에서의 관심지수는 단기간 동안 사용자들의 관심도를 정확히 파악하기 어렵지만, 장시간 동안 많은 관심을 유도하는 토픽이 무엇인지 알 수 있다.

3.1.3 핫토픽 사전 탐지 단계

핫토픽 선정 단계에서 추출한 핫토픽의 결과를 종속변수로 사용하여 SAS Enterprise Miner Workstation 12.3을 통해 로짓, 의사결정나무, SVM 등의 기법으로 학습시켜 핫토픽 사전 탐지를 시행한다. 각각의 기법을 이용하여 성과를 평가한다. 먼저, 5개 입력변수로 월별(t+1) 및 분기별(t+3)로 적용한 후 핫토픽 사전 탐지에 대한 결과를 비교한다. 또한, 월별 및 분기별에 발생된 모든 토픽을 고려하여 동일한 핫토픽을 선정한 후 관심지수를 계산한다. 최종적으로 핫토픽 선정 단계에서의 관심지수와 핫토픽 사전 탐지 단계에서의 관심지수에 대한 결과를 비교하여 사용자들에게 추천해 줄 수 있는 핫토픽을 선정하게 된다.

4. 실험 및 결과분석

4.1 데이터 수집

본 연구에서는 중국의 시나 웹사이트에서 온라인 주식 포럼 데이터를 추출하였다.

〈Table 2〉 Categories in Online Sina Stock Forum

A股 (stock A)	理財 (investment techniques)	主題 (subject)	B股 (stock B)	H股 (stock A)	基金 (fund)
期貨 (future)	權証 (warent)	証券公司 (securities company)	大陸指數 (mainland index)	港股指數 (HongKong index)	債券 (bond)
經濟學人 (economist)	財經視点 (financial view)	基金公司 (fund conpany)	財智大贏家 (financial expert)	美股100 (US index 100)	財經明星 (finance star)
商學院 (business school)	股指期貨 (index future)	農産品 (agricultural commodities)			

시나는 현재 블로그, 마이크로블로그 및 포럼 등 다양한 분야로 발전하여 사용되고 있으며, 약 2.49억 명에 달하는 회원을 보유하고 있다. 시나 주식 포럼에서 제공되는 총 21개의 카테고리는 <Table 2>와 같이 나타낼 수 있다.

본 연구에서는 21개의 카테고리들 중 '주제'에 해당하는 포럼을 선택하여 <Table 3>과 같이 총 144개에 해당하는 토픽들이 포함되어 있는 것을 확인하였다.

〈Table 3〉 The Parts Included in the 'Subject' Category

new energy	olympic	insurance
real estate	long-term investment	shot-term investment
dark horse	value investing	technology exchange
stock analysis	low priced stock	new stock
....	....	....

토픽에 해당하는 포스팅을 수집하기 위해 자동 수집 프로그램인 빠쫘위(八爪鱼)를 사용하여 데이터를 크롤링한 후 포스팅이 없거나 관련성

이 낮은 토픽을 제외시켜 2013년 3월부터 2015년 2월까지의 데이터를 수집한 결과 총 68개의 토픽과 21,141개의 포스팅을 추출할 수 있었다. 또한, 24개월간 포스팅에 대한 조회 수는 최고 155,903번에서 최저 1,180번의 조회로 나타났으며, 주식에 대한 전문가의 해석이 내포되어 있는 토픽이 시간이 지남에 따라 지속적으로 활발한 토픽임을 알 수 있으며, 월별 토픽수와 포스팅의 수는 <Table 4>와 같이 나타났다.

〈Table 4〉 The Number of Postings and Monthly Topics

Date	Topic No.	Posting No.	Date	Topic No.	Posting No.
2013.03	44	786	2014.03	48	662
2013.04	32	573	2014.04	54	972
2013.05	36	476	2014.05	57	1,187
2013.06	30	586	2014.06	59	1,216
2013.07	34	568	2014.07	55	1,226
2013.08	40	488	2014.08	56	1,409
2013.09	37	443	2014.09	57	1,417
2013.10	48	572	2014.10	60	1,026
2013.11	44	685	2014.11	61	1,061
2013.12	47	734	2014.12	60	1,166
2014.01	50	614	2015.01	62	1,610
2014.02	47	587	2015.02	63	1,077

4.2 핫토픽 선정

4.2.1 감성값 계산 및 극성분류

HowNet의 감성분석을 위한 용어 사전에 수록된 어휘의 극성을 5가지로 분류하여 감성값을 계산한다. 각 어휘의 극성을 구분하기 위해 단어의 표현 정도에 따라 1점, 3점, 5점, 7점, 9점으로 설정하고, 개별적으로 분리하여 문서와의 매칭을 통해 감성값을 산출한다. 중국어의 경우 문장을 분할시켜야 함으로 프로그램 'R'을 이용한다. 예를 들어 문서 내에 있는 문장이 '我爱釜山'로 작성되어 있다면 "我" "爱" "釜山"로 분할한다. 따라서, 하나의 포스팅에 대한 감성값은 포스팅에 내포되어 있는 단어의 감성점수의 합으로 구성되어 감성값을 산출하게 되는 것이다.

〈Table 5〉 Class of HowNet Sentiment Dictionary

稍微	較	很	超	极其
insufficiently	more	very	over	extreme
조금	더욱	매우	초과	극한
1	3	5	7	9

문서의 극성분류는 용어 사전을 이용하여 긍정/부정에 대한 감성 단어 사전과 긍정/부정 평가 단어 사전으로 분류하고, 문서와 매칭시켜 긍정 단어와 부정 단어를 찾아 문서의 극성을 분류하게 된다. 문서 내에 있는 긍정 단어의 수가 부정단어의 수보다 많으면 긍정문서, 적으면 부정문서, 같으면 중립문서로 분류되며 최종적으로 핫토픽을 탐지할 때 하나의 토픽 내에 포함되어 있는 긍정 및 부정 포스팅의 수를 확인한다.

〈Table 6〉 A Number of Words in Polarity Dictionary

Emotional polarity dictionary	No. of words
positive emotion terms	836
negative emotion terms	1,254
positively valued terms	3,730
negatively valued terms	3,116

4.2.2 군집분석 결과

4.2.2.1 k-means 및 SOM 군집결과

5개의 입력변수들을 사용하여 월별로 나누고 최대 k값을 구한 후 시행착오 방법을 통하여 6,

11, 16로 정하여 k-means 군집을 시행하였다. 24개월에 해당되는 모든 토픽들의 수는 총 1,181개로 나타났다. k가 6일 경우 핫토픽의 수는 144개, k가 11일 경우 264개, k가 16일 경우 384개로 집계되었으며 <Table 7>과 같이 나타나는 것을 확인할 수 있었다.

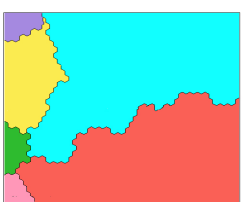
〈Table 7〉 A Number of Hot Topic Using K-means

	k=6	k=11	k=16
0	1,037	917	797
1	144	264	384
Total	1,181	1,181	1,181

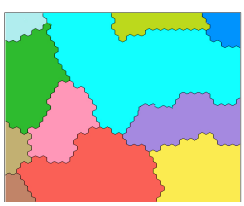
0: Normal topic, 1: Hot topic

예를 들어, 2013년3월 데이터에서는 사용자들의 개별적 의견이 담긴 토픽들이 주로 핫토픽이 되었지만, 2015년 2월 데이터에서는 전문가의 의견이 담긴 토픽들이 주로 핫토픽이 된 것을 확인할 수 있었다. 사용자들이 주식에 대한 관심이 점차 증가하기 시작하면서 전문가의 의견을 중점으로 많은 정보를 이용하기 때문에 이러한 토픽들이 핫토픽이 될 것이라 판단된다.

SOM(Self-Organizing Map) 기법으로 실험하기 위해 Viscovery SOMine 6를 이용하고, k-means 기법과 동일한 입력변수를 사용하였다. 2013년 3월의 데이터를 통해 군집의 수를 6개와 11개로 설정한 후 군집에 대한 결과로 <Figure 2>와 <Figure 3>과 같이 나타나는 것을 확인할 수 있었다.



〈Figure 2〉 Results of SOM(6 Clusters)



〈Figure 3〉 Results of SOM(11 Clusters)

<Figure 2>와 <Figure 3>에서처럼 토픽의 빈도수로 표현되어 k-means기법에서의 군집에 대한 결과와 비슷한 양상을 보이고 있는 것을 확인할 수 있었다. 또한, 다른 월별 데이터에서도 실험을 통해 비슷한 결과가 나타나는 것을 확인할 수 있었다. 본 연구에서는 k-means 기법에서 핫토픽을 찾아내는 과정이 더 용이하고 좋은 해석력을 가지고 있다고 판단하여 k-means 기법을 이용한 결과를 이용하였다.

4.3 핫토픽 사전 탐지 모형

핫토픽을 탐지를 위해 전체 기간에 대해 동일하게 나타난 토픽은 총 68개 중 17개의 토픽으로 나타나는 것을 확인할 수 있었다. 24개월에 대한 핫토픽을 사전 탐지하기 위해 월별(t+1)과 분기별(t+3)로 k가 6, 11, 16의 값에 따라 로짓, 의사결정나무, SVM 기법을 이용하여 학습용 데이터와 검증용 데이터를 8:2로 분할하여 실험을 실시하였다.

4.3.1 로짓을 이용한 핫토픽 사전 탐지 모형 결과

로짓을 이용한 핫토픽 사전 탐지 모형에 대해 월별 실험결과 k가 6일 경우 84.62%, k가 11일 경우 75.95%, k가 16일 경우 62.03%의 예측정확

〈Table 8〉 Results of Hot Topic Detection Model Using Logit

	Monthly		Quarterly	
	Training Set	Testing Set	Training Set	Testing Set
k=6	87.23%	84.62%	85.52%	86.46%
k=11	73.40%	75.95%	78.95%	77.78%
k=16	64.11%	62.03%	65.73%	67.57%

성을 나타내고 있었다. 분기별 실험결과 k가 6일 때 86.46%, k=11일 때 77.78%, k가 16일 때 67.57%의 예측정확성을 나타내고 있었다.

4.3.2 의사결정나무를 이용한 핫토픽 사전 탐지 모형 결과

의사결정나무를 이용한 핫토픽 사전 탐지 모형에 대한 결과를 알아보기 위해 CHAID, CART와 C4.5를 사용하여 실험을 실시하였다. 3가지 기법을 이용하여 실험한 결과 CHAID를 이용한 핫토픽 사전 탐지 모형의 실험 결과가 다른 모형들에 비해 월별과 분기별 모두 좋은 예측성과를 나타내고 있었다.

〈Table 9〉 Results of Hot Topic Detection Model Using CHIAID

	Monthly		Quarterly	
	Training Set	Testing Set	Training Set	Testing Set
k=6	88.18%	84.62%	87.64%	82.44%
k=11	74.04%	75.95%	79.65%	75.00%
k=16	66.35%	62.03%	66.44%	58.11%

CHIAID를 이용한 핫토픽 사전 탐지 모형의 실험결과 월별로 핫토픽을 사전 탐지할 경우 k의 값이 6일 때 84.62%, k가 11일 때 75.95%, k가 16일 때 62.03%의 예측 정확성으로 나타났으며,

〈Table 10〉 Results of Hot Topic Detection Model Using CART

	Monthly		Quarterly	
	Training Set	Testing Set	Training Set	Testing Set
k=6	90.10%	76.93%	89.40%	82.44%
k=11	79.81%	67.09%	86.32%	72.23%
k=16	77.57%	53.17%	80.22%	60.82%

분기별로 핫토픽을 사전 탐지할 경우 k가 6일 때 82.44%, k가 11일 때 75.00%, k가 16일 때 58.11%의 예측정확성으로 나타났다.

CART를 이용한 핫토픽 사전 탐지 모형의 실험결과 월별로 핫토픽을 사전 탐지할 경우 k의 값이 6일 때 76.93%, k가 11일 때 67.09%, k가 16일 때 53.17%의 예측 정확성으로 나타났으며, 분기별로 핫토픽을 사전 탐지할 경우 k가 6일 때 82.44%, k가 11일 때 72.23%, k가 16일 때 60.82%의 예측정확성으로 나타났다.

〈Table 11〉 Results of Hot Topic Detection Model Using C4.5

	Monthly		Quarterly	
	Training Set	Testing Set	Training Set	Testing Set
k=6	90.42%	76.93%	89.40%	81.09%
k=11	78.21%	72.16%	87.02%	61.12%
k=16	76.29%	53.17%	80.22%	59.46%

C4.5를 이용한 핫토픽 사전 탐지 모형의 실험결과 월별로 핫토픽을 사전 탐지할 경우 k가 6일 때 76.93%, k가 11일 때 72.16%, k가 16일 때 53.17%의 예측 정확성으로 나타났으며, 분기별로 핫토픽을 사전 탐지할 경우 k가 6일 때 k가 11일 때 81.09%, 61.12%, k가 16일 때 59.46%의 예측정확성으로 나타났다.

4.3.3 SVM을 이용한 핫토픽 사전 탐지 모형 결과

SVM을 이용한 핫토픽 사전 탐지 모형에 사용된 커널함수는 가우시안RBF(Gaussian Radial Basis Function kernel)를 이용하였으며, 파라미터 C는 C = {1, 25, 50, 75, 100}으로 설정하였고, σ 값은 σ = {0.25, 0.5, 1, 5, 10}으로 설정하여 격

자 탐색(grid search)을 통해 최적의 파라미터를 선정하여 실험결과를 도출하였다.

실험결과 월별로 핫토픽을 사전 탐지할 경우 k가 6일 때 C=50, σ=5에서 84.62%, k가 11일 때 C=50, σ=0.5에서 73.42%, k가 16일 때 C=75, σ=1에서 63.30%의 예측정확성으로 나타났으며, 분기별로 핫토픽을 사전 탐지할 경우 k가 6일 때 C=50, σ=1에서 86.49%, k가 11일 때 C=50, σ=0.5에서 77.78%, k가 16일 때 C=25, σ=0.5에서 74.33%의 예측정확성으로 나타났다.

〈Table 12〉 Results of Hot Topic Detection Model Using SVM

	Monthly		Quarterly	
	Training Set	Testing Set	Training Set	Testing Set
k=6	97.77%	84.62%	86.93%	86.49%
k=11	74.36%	73.42%	78.25%	77.78%
k=16	68.92%	63.30%	63.61%	74.33%

4.4 관심지수 결과 분석

관심지수는 단시간 동안 사용자들의 관심도를 정확히 파악하기 어렵지만, 장시간 동안 많은 관심을 유도하는 토픽이 무엇인지 알 수 있다.

핫토픽 선정 단계에서의 관심지수와 핫토픽 사전 탐지 단계에서의 관심지수를 이용하여 월별 및 분기별에 발생된 모든 토픽을 고려한 후 24개월에 해당하는 핫토픽의 관심지수를 계산하였다. 핫토픽 선정 단계와 핫토픽 사전 탐지 단계에서 추출된 핫토픽을 <Table 13>~<Table 15>와 같이 요약하였으며, 관심지수가 높은 핫토픽을 기준으로 순위화하여 상위 5개에 해당하는 핫토픽을 나타내었다.

<Table 13>에서처럼 핫토픽 선정 단계에서 k가 6일 때의 핫토픽과 로짓을 이용한 핫토픽 사전 탐지 모형과 의사결정나무를 이용한 핫토픽 사전 탐지 모형의 결과에서 3개(60번, 68번, 7번)의 핫토픽이 일치하는 것을 확인할 수 있었고, 로짓을 이용한 핫토픽 사전 탐지 모형과 의사결정나무를 이용한 핫토픽 사전 탐지 모형이 일치하는 것을 확인할 수 있었다.

〈Table 13〉 Compare of Hot Topic Interest Index(k=6)

Hot Topic Selection		Hot Topic Detection					
K-means		Logit		Decision Tree		SVM	
A	B	A	B	A	B	A	B
60	15	60	11	60	14	60	14
68	11	68	6	68	6	68	10
7	7	7	2	7	2	63	5
8	3	34	2	34	2	34	5
14	2	6	1	6	1	21	4
A: hot topic, B: interest index							

〈Table 14〉 Compare of Hot Topic Interest Index(k=11)

Hot Topic Selection		Hot Topic Detection					
K-means		Logit		Decision Tree		SVM	
A	B	A	B	A	B	A	B
60	22	60	20	60	17	60	18
68	17	68	8	7	6	7	5
7	12	7	6	34	3	68	4
34	11	34	3	56	1	34	3
44	10	44	3	68	0	44	1
A: hot topic, B: interest index							

<Table 14>와 같이 핫토픽 선정 단계에서 k가 11일 때의 핫토픽과 로짓을 이용한 핫토픽 사전 탐지 모형과 SVM을 이용한 핫토픽 사전 탐지

모형에서 동일한 핫토픽으로 나타나는 것을 확인할 수 있었다. 의사결정나무를 이용한 핫토픽 사전 탐지 모형의 결과에서는 56번을 제외하고 다른 모형과의 일치성을 나타내고 있었다.

핫토픽 선정단계에서 k가 16일 때의 핫토픽과 로짓을 이용한 핫토픽 사전 탐지 모형, 의사결정나무를 이용한 사전 탐지 모형, SVM을 이용한 사전 탐지 모형 모두에서 3개(60번, 68번, 7번)의 핫토픽이 일치하는 것을 확인할 수 있었고, k가 6, 11일 때의 결과보다 높은 관심지수를 나타내고 있었다.

〈Table 15〉 Compare of Hot Topics Interest Index(k=16)

Hot Topic Selection		Hot Topic Detection					
K-means		Logit		Decision Tree		SVM	
A	B	A	B	A	B	A	B
60	23	60	23	60	22	60	23
68	16	68	12	7	6	7	7
7	15	7	11	8	4	68	5
25	14	56	6	21	2	34	3
48	11	48	5	34	2	44	3
A: hot topic, B: interest index							

로짓을 이용한 핫토픽 사전 탐지 모형에서 예측성과가 높게 나타났지만, 관심지수를 이용하여 핫토픽을 비교하였을 때 단기간이 아닌 장기간 지속적으로 사용자들로부터 관심을 유도한 핫토픽의 경우 k가 11일 때의 성과가 가장 우수하다는 것을 알 수 있었다. 즉, 관심지수를 기준으로 순위화하였을 때 상위 5개의 핫토픽이 일치하여 사용자들에게 가장 좋은 핫토픽을 추천할 수 있게 된다는 것이다. 따라서 핫토픽 선정단계에서의 관심지수와 핫토픽 사전 탐지 단계에서의 관심지수를 비교한 결과 효과적인 핫토픽을 찾을 수 있었다.

5. 결론 및 향후 연구 방향

본 연구에서는 온라인 시나 주식 포럼 데이터를 이용하여 감성단어를 바탕으로 감성분석에 대한 연구를 진행하였다. 또한, 어휘기반 접근방법인 감성사전의 사용과 기계학습 접근기법을 이용하여 통합 접근방법의 연구로 온라인 포럼에서의 다양한 토픽들을 통해 핫토픽을 선정하고 탐지하였다. 2014년도부터 중국의 주식시장이 활발해지면서 온라인 주식 포럼에서도 많은 영향을 미치게 되어 전문가의 의견을 담은 토픽들이 활성화되기 시작하였다. 즉, 전문가들의 의견을 담고 있는 포스팅과 일반 사용자들이 작성한 포스팅의 수가 많을수록 핫토픽이 된다는 것을 알 수 있다. 본 실험을 통해 '주제'에 해당하는 포럼에서 총 144개에 해당하는 토픽들이 포함되어 있지만, 사용자들이 많은 관심을 가지는 토픽은 68개로 집계되었다. 24개월간의 포스팅에 대한 조회 수를 확인한 결과 최고 155,903번에서 최저 1,180번의 조회를 나타내고 있었으며, 많은 사용자들이 관심을 가지고 이용하고 있다는 것을 알 수 있었다. 또한, 관심지수가 높은 핫토픽을 기준으로 순위화하여 상위 5개에 해당하는 핫토픽을 사용자들에게 추천함으로써 다른 사용자들에게 유용한 정보를 제공하고 보다 질 좋은 토픽을 추천해 줄 뿐만 아니라 투자자들에게도 많은 도움을 줄 수 있을 것이라 기대된다.

본 연구를 통해 핫토픽을 추천함으로써 최신 동향 및 주식 투자를 원하는 사용자들에게 좋은 정보가 될 것이며 가장 효율적이고 유용한 정보를 쉽게 활용할 수 있다는 장점을 가질 수 있다. 본 연구에서는 2013년 3월부터 2015년 2월까지

수집된 데이터를 사용하였기 때문에 최신 정보를 추가한다면 데이터를 빠르고 정확하게 분석할 수 있는 실시간형 서비스가 될 수 있다. 또한, 소셜 네트워크 서비스를 이용하는 사용자들의 관심 주제나 키워드를 선정하여 핫토픽을 선정한다면 사용자의 니즈를 보다 쉽게 파악할 수 있을 것으로 기대되며, 리뷰를 기반으로 분석이 이루어지는 감성분석의 연구에서 감성, 감정 및 극성을 분류하고 분석할 수 있는 기초자료가 될 것이라 생각된다.

본 연구에서는 24개월에 해당하는 온라인 시나 주식 포럼 데이터를 이용하여 월별 및 분기별로만 핫토픽을 탐지하였지만 향후 연구에서는 좀 더 세부적으로 주별 단위의 연구가 진행되어야 할 것이다. 또한, 쇼핑정보 공유사이트, 브랜드 포럼 등 다양한 포럼 데이터를 활용하여 마케팅 측면에서 판매를 촉진시킬 수 있는 연구가 진행되어야 할 것이다.

참고문헌(References)

An, J. and H. Kim, "Building a Korean Sentiment Lexicon Using Collective Intelligence," *Journal of Intelligence and Information Systems*, Vol.21, No.2(2015), 49～67.

Baccianella, S., A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *Proceedings of the Seventh Conference on International Language Resources and Evaluation,* Vol.10(2010), 2200～2204.

Bollen, J., H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, Vol.2, No.1(2011), 1～8.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, 2008.

Burges, C. J., "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, Vol.2, No.2(1998), 121～167.

Chen, L., L. Qi, and F. Wang, "Comparison of feature-level learning methods for mining online consumer reviews," *Expert Systems with Applications*, Vol.39(2012), 9588～9601.

Fung, G. P. C., J. X. Yu, and W. Lam, "Stock prediction: Integrating text mining approach using real-time news," *Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering*, (2003), 395～402.

Hartigan, J. A., *Clustering Algorithms*. John Wiley & Sons, Inc., 1975.

Hong, T. and E. Kim, "Predicting the Response of Segmented Customers for the Promotion Using Data Mining," *Information Systems Review*, Vol.12, No.2(2010), 75～88.

Hu, M. and B. Liu, "Mining Opinion Features in Customer Reviews," *Proceedings of the 19th national conference on Artificial intelligence,* (2004), 755～760.

Huang, C. J., J. J. Liao, D. X. Yang, T. Y. Chang, and Y. C. Luo, "Realization of a news dissemination agent based on weighted association rules and text mining techniques," *Expert Systems with Applications*, Vol.37, No.9(2010), 6409～6413.

Hu, N., I. Bose, N. S. Koh, and L. Liu, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decision Support Systems*, Vol.52, No.3 (2012), 674～684.

Jin, F., N. Self, P. Saraf, P. Butler, W. Wang, and

N. Ramakrishnan, "Forex-foreteller: Currency trend modeling using news articles," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, (2013), 1470～1473.

Jin, Y., J. Kim, and J. Kim, "Product Community Anlaysis Using Opinion Mining and Network Anlysis: Movie Performance Prediction Case," *Journal of Intelligence and Information Systems*, Vol.20, No.1(2014), 49～165.

Kass, G., "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, Vol.29(1980), 119～127.

Kim, Y. M., S. J. Jeong, and S. J. Lee, " A Study on the Stock Market Prediction Based on Sentiment Analysis of Social Media," *Entrue Journal of Information Technology*, Vol.13, No.3(2014), 59～70.

Li, N. and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, Vol.48, No.2(2010), 354～368.

Liu, B., "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, Vol.5, No.1(2012), 1～167.

Maks, I. and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," *Decision Support Systems*, Vol.53, No.4(2012), 680～688.

Martín-Valdivia, M. T., E. Martínez-Cámara, J. M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications*, Vol.40, No.10(2013), 3934～3942.

Medhat, W., A. Hassan, and H. Korashy, "Sentiment analysis algorithms and application: A survey," *Ain Shams Engineering Journal*, Vol.5(2014), 1093～1113.

Oh, S.-H. and S.-J. Kang, "Movie Retrieval System by Analyzing Sentimental Keyword from User's Movie Reviews," *Journal of the Korea Academia-Industrial*, Vol.14, No.3(2013), 1422～1427.

Pang, B. and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2 (2008), 1～135.

Pang, B., L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol.10(2002), 79～86.

Park, H. and K. H. Cho, "CHAID Algorithm by Cubebased Proportional Sampling," *Journal of Korean Data & Information Science Society*, Vol.15, No.4(2004), 803～816.

Quinlan, J. R., "Induction of Decision Trees," *Machine Learning*, Vol.1, No.1(1986), 81～106.

Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California, 1993.

Schumaker, R. P., Y. Zhang, C. N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decision Support Systems,* Vol.53, No.3(2012), 458～464.

Tan, S. and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with Applications*, Vol.34, No.4(2008), 2622～2629.

Turney, P. D. and M. L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions*

on Information Systems(TOIS), Vol.21, No.4 (2003), 315～346.

Vapnik, V., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

Wang, G., J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision Support Systems*, Vol.57(2014), 77～93.

Yu, E., Y. Kim, N. Kim, and S. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems*, Vol.19, No.1(2013), 95～110.

Zhang, C., D. Zeng, J. Li, F. Y. Wang, and W. Zuo, "Sentiment analysis of Chinese documents: From sentence to document level," *Journal of the American Society for Information Science and Technology*, Vol.60, No.12(2009), 247 4～2487.

Abstract

Development of Sentiment Analysis Model for the hot topic detection of online stock forums

Taeho Hong* · Taewon Lee** · Jingjing Li***

Document classification based on emotional polarity has become a welcomed emerging task owing to the great explosion of data on the Web. In the big data age, there are too many information sources to refer to when making decisions. For example, when considering travel to a city, a person may search reviews from a search engine such as Google or social networking services (SNSs) such as blogs, Twitter, and Facebook. The emotional polarity of positive and negative reviews helps a user decide on whether or not to make a trip. Sentiment analysis of customer reviews has become an important research topic as datamining technology is widely accepted for text mining of the Web. Sentiment analysis has been used to classify documents through machine learning techniques, such as the decision tree, neural networks, and support vector machines (SVMs). is used to determine the attitude, position, and sensibility of people who write articles about various topics that are published on the Web. Regardless of the polarity of customer reviews, emotional reviews are very helpful materials for analyzing the opinions of customers through their reviews. Sentiment analysis helps with understanding what customers really want instantly through the help of automated text mining techniques. Sensitivity analysis utilizes text mining techniques on text on the Web to extract subjective information in the text for text analysis. Sensitivity analysis is utilized to determine the attitudes or positions of the person who wrote the article and presented their opinion about a particular topic.

In this study, we developed a model that selects a hot topic from user posts at China's online stock forum by using the k-means algorithm and self-organizing map (SOM). In addition, we developed a detecting model to predict a hot topic by using machine learning techniques such as logit, the decision tree, and SVM. We employed sensitivity analysis to develop our model for the selection and detection of hot topics from China's online stock forum. The sensitivity analysis calculates a sentimental value from a document based on contrast and classification according to the polarity sentimental dictionary (positive

* College of Business Administration, Pusan National University
** Corresponding Author: Taewon Lee
Institute of China Studies, Pusan National University
2, Busangaehag-ro 63 Beon-gil, Geumjeong-gu, Busan 609-735, Korea
Tel: +82-51-510-3988, Fax: +82-51-510-3989, E-mail: wantty@pusan.ac.kr
*** Institute of Chinese Studies, Pusan National University

or negative).

The online stock forum was an attractive site because of its information about stock investment. Users post numerous texts about stock movement by analyzing the market according to government policy announcements, market reports, reports from research institutes on the economy, and even rumors. We divided the online forum's topics into 21 categories to utilize sentiment analysis. One hundred forty-four topics were selected among 21 categories at online forums about stock. The posts were crawled to build a positive and negative text database. We ultimately obtained 21,141 posts on 88 topics by preprocessing the text from March 2013 to February 2015. The interest index was defined to select the hot topics, and the k-means algorithm and SOM presented equivalent results with this data. We developed a decision tree model to detect hot topics with three algorithms: CHAID, CART, and C4.5. The results of CHAID were subpar compared to the others. We also employed SVM to detect the hot topics from negative data. The SVM models were trained with the radial basis function (RBF) kernel function by a grid search to detect the hot topics.

The detection of hot topics by using sentiment analysis provides the latest trends and hot topics in the stock forum for investors so that they no longer need to search the vast amounts of information on the Web. Our proposed model is also helpful to rapidly determine customers' signals or attitudes towards government policy and firms' products and services.

Key Words : Sentiment Analysis, Opinion Mining, SVM, Hot topic, Online forums.

Received : September 9, 2015 Revised : March 15, 2016 Accepted : March 16, 2016
Publication Type : Regular Paper Corresponding Author : Taewon Lee

저 자 소 개

홍 태 호

현재 부산대학교 경영대학 교수로 재직하고 있다. KAIST에서 산업공학사를 취득하였고 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 딜로이트 컨설팅에서 컨설턴트로 재직했으며, 주요 관심분야는 비즈니스 애널리틱스, 데이터 마이닝, 오피니언 마이닝, 소셜네트워크 분석, 지식 발굴 및 경영 등이다.

이 태 원

현재 부산대학교 중국연구소에서 전임연구원으로 재직하고 있다. 동국대학교 학사를 취득하였고, 영남대학교에서 공학석사를 취득하였으며, 부산대학교에서 경영정보시스템을 전공하여 경영학박사를 취득하였다. 주요 연구분야는 빅데이터, 데이터마이닝, 오피니언 마이닝, CRM, 소셜네트워크 분석 등이다.

리 징 징

현재 부산대학교 중국연구소에서 전임연구원으로 재직하고 있다. 하얼빈사범대학교 학사를 취득하였고 부산대학교에서 경영정보시스템을 전공하여 석사를 취득하였다. 주요 연구분야는 오피니언마이닝, 소셜네트워크분석, 비즈니스 애널리틱스 등이다.