Journal of Korean Institute of Intelligent Systems, Vol. 24, No. 3, June 2014, pp. 292-297
http://dx.doi.org/10.5391/JKIIS.2014.24.3.292

트위터를 활용한 감성 기반의 영화 유사도 측정

Measuring Similarity Between Movies Based on Sentiment of Tweets

김경민 · 김동윤 · 이지형†

Kyoungmin Kim, Dong-Yun Kim, and Jee-Hyong Lee†

성균관대학교 정보통신대학

College of Information and Communication Engineering, Sungkyunkwan University

요 약

최근 소셜 네트워크 서비스가 보편화되면서, 이를 활용하여 사람들의 의견이나 감성 등을 파악하기 위한 감성분석 연구가 다양한 분야 진행되고 있다. 기존의 영화 관련 연구의 경우, 대부분이 영화평에 대해 단순 긍/부정으로 감성분석을 하여, 영화에 대한 선호도를 파악하는 데 그쳤다. 사람의 감성은 단순 긍/부정이 아닌 다양한 감성으로 분류될 수 있는데 반해, 이분법적 감성분석은 영화의 평점 정보에서 손쉽게 얻을 수 있는 선호도와 유사한 분석을 하는데 그친다. 따라서 영화의 평점보다 다양하고 유용한 정보를 얻기 위해서는, 영화 리뷰를 세분화된 감성으로 분석하여 영화에 대해 느낀 감성을 다양한 기준으로 분류할 필요가 있다. 본 논문에서는 Thayer 모델을 기반으로 감성 분류 기준을 세우고, 수집한 영화 관련 트윗을 이용하여 각 영화에 대해 대중이 느끼는 감성을 분석한다. 분석된 영화에 대한 감성 비율을 유클리드거리, 코사인유사도, 피어슨 상관계수를 이용하여 영화간의 유사도를 측정하였다. IMDB에서 제공하는 유사 영화 정보를 바탕으로 본 논문에서 제안하는 방식의 유용성을 검증하였다.

키워드 : 데이터 마이닝, 감성분석, 트위터, 유사도 측정

Abstract

As a Social Network Service (SNS) has become an integral part of our everyday lives, millions of users can express their opinion and share information regardless of time and place. Hence sentiment analysis using micro-blogs has been studied in various field to know people's opinion on particular topics. Most of previous researches on movie reviews consider only positive and negative sentiment and use it to predict movie rating. As people feel not only positive and negative but also various emotion, the sentiment that people feel while watching a movie need to be classified in more detail to extract more information than personal preference. We measure sentiment distributions of each movie from tweets according to the Thayer's model. Then, we find similar movies by calculating similarity between each sentiment distributions. Through the experiments, we verify that our method using micro-blogs performs better than using only genre information of movies.

Key Words : Data Mining, Sentiment Analysis, Twitter, Similarity Measurement

1. 서 론

최근 소셜 네트워크 서비스(Social Network Service; SNS)가 보편화되면서, 기존 오프라인 방식과 달리 시간과 공간의 제약 없이 자신의 개인 생활 및 영화, 제품 등 다양한 분야에 대한 의견을 자유롭게 공유할 수 있게 되면서, 이를 통한 정보 생성과 공유가 활발히 이루어지고 있다.

대표적인 SNS인 트위터는 2006년 3월에 서비스를 도입 이후, 현재 전 세계적으로 사용자가 계속 증가하는 추세이며, 2012년 2월 기준 트위터에 등록된 사용자 수는 4억 6천 5백만명 이상으로, 하루 평균 1억 7천 5백만개 이상의 트윗이 발행된다.[1] 이처럼 계속적으로 축적되는 방대한 양의 데이터를 활용하기 위해 데이터로부터 의미 있는 정보를 추출해 내는 기술 중 하나인 데이터마이닝의 필요성이 증대되고 있다. 특히, 데이터에서 어떤 사안이나 인물, 이슈에 대한

접수일자: 2013년 9월 1일
심사(수정)일자: 2013년 9월 7일
게재확정일자 : 2013년 12월 18일
† 교신저자(Corresponding author)
본 연구는 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업의 연구결과입니다. (NRF-2012R1A1A2008062)

또한, 본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였습니다. (10041244, 스마트TV 2.0 소프트웨어 플랫폼) 연구비 지원에 감사드립니다.

1) http://infographiclabs.com/news/twitter-2012

사람들의 의견이나 감성 등을 파악하는 감성분석(sentiment analysis)은 다양한 분야에서 활용되고 있다[1]. 대표적인 예로 유니레버 도브 사의 경우, 감성분석을 활용하여 자사의 캠페인에 대한 소비자의 반응을 파악하여 더 나은 제품을 제공함으로써 소비자의 만족도를 높인 바 있다[2].

대부분의 오피니언 마이닝 연구들은 대중의 생각이 긍정적인지 부정적인지를 판단하여 특정 대상에 대한 대중의 선호도를 예측한다. 영화와 관련된 감성분석 연구의 경우, 주로 영화평의 감성분석을 통해 영화에 대한 대중의 선호도 또는 영화에 대한 평점이나 흥행을 예측하는 연구가 활발하게 진행되고 있다[3-6].

사람의 감성은 단순 긍/부정이 아닌 슬픔, 무서움, 기쁨 등 다양한 감성으로 분류될 수 있는데 반해, 영화평을 단순 긍/부정으로 분류하는 이분법적 감성분석은 영화의 평점 정보에서 손쉽게 얻을 수 있는 선호도와 유사한 분석을 하는데 그친다. 따라서 영화의 선호도 외에 다양하고 유용한 정보를 얻기 위해서는, 영화에 대해 느낀 감성을 다양한 기준으로 분류함으로써 영화에 대한 세분화된 감성분석이 필요하다. 이는, 영화 장르, 줄거리 등과 같은 영화에 대한 메타데이터와 함께 사용자에게 영화를 추천하는 추천시스템 등 다양한 분야에서 활용될 수 있다.

본 연구진은 서로 다른 두 영화가 유사한지 판단할 때 고려하는 기준에 대해 20, 30대 남녀 124명을 대상으로 설문조사를 시행하였다. 설문조사는 구글독스(Google Docs)를 이용하여 온라인으로 진행되었으며 설문조사 결과, 영화가 유사한지 판단할 때 고려하는 기준으로 설문 응답자의 65%가 영화가 주는 감성을 고려한다고 응답하였다. 이는 유사한 영화를 찾는데 있어, 소셜 미디어를 통해 영화를 보고 느끼는 대중의 감성분석이 유용한 정보가 될 수 있음을 보여주는 것뿐만 아니라, 대중의 영화에 대한 감성이 추천시스템의 만족도를 높일 수 있는 요인이 될 수 있음을 보여준다.

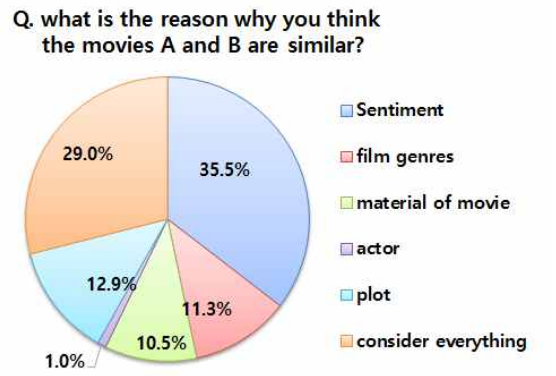

그림 1. 유사 영화의 이유에 대한 설문조사 결과
Fig. 1. The Result of a survey about the reason for choosing similar movies

본 논문에서는 Thayer 모델[7]을 기반으로 사람이 느끼는 감성의 분류 기준을 정의하고, 영화와 관련된 트윗의 감성분석을 통해 영화간의 감성 유사도와 영화의 장르간의 유사도를 이용하여 영화간의 유사도를 측정하는 방식을 제안한다. 다수의 트위터 사용자들이 작성한 트윗을 이용한 감성분석을 통해 영화간의 유사도를 측정함으로써, 영화에 대해 대중이 느끼는 다양한 감성 및 대중의 선호도를 반영한 영화간의 유사도를 구할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해서 설명하고, 3장에서는 본 논문에서 제안하는 방법에 대해 기술한다. 4장에서 실험결과를 분석을 하고, 마지막으로 5장에서 결론 및 향후 연구에 대해서 논의한다.

2. 관련연구

Oghina, Andrei, 외 3명의 연구에서는 영화의 평점을 예측하기 위해 영화의 메타데이터 및 트윗과 같은 소셜 미디어 정보를 활용하였다[3]. 해당 연구에서는 트윗에서 영화관련 어휘가 언급되는 횟수, 영화에 대한 트윗들의 긍/부정의 비율, 유투브에서 해당 영화에 관련된 영상의 like/dislike 수 등을 이용하여 영화의 평점을 예측하였다. 소셜 미디어 정보를 활용한 경우와, 영화의 메타 데이터만을 이용하여 평점예측 성능의 비교 결과, 소셜 미디어 정보를 활용한 경우의 예상 평점이 실제 값과 유사한 결과를 보였다.
이 연구를 통해서 소셜 미디어 정보 활용의 가치를 확인할 수 있었으나, 영화에 대한 트윗이 나타나는 감성을 긍/부정으로만 한정지어, 선호도를 예측하는데 유용한 정보가 될 수 있으나, 유사한 영화를 찾는데 적합하지 못하다.

Jana Eggink 외 1명의 연구에서는 TV프로그램에서 제공하는 장르와 같은 기본적인 정보 외에 기계 학습을 통해 자동으로 얻은 감성 정보를 활용하는 방법을 제안하였다[4]. 사람이 느끼는 감성을 분류하기 위해 Osgood의 유의어 사전을 활용하여 sad/happy, serious/humorous, exciting/relaxing, interesting/boring, slow/fast-paced, light-hearted/dark 의 6가지 감성의 짝에 interesting/boring을 추가하여 총 7가지의 감성의 짝을 기준으로 분류하였다. 그리고 TV프로그램의 감성 정보와 장르 간의 유의미한 상관관계 분석을 통해 7가지 감성의 짝 중 유사한 것을 차원 축소를 통해 2차원의 공간에 감성을 매핑시켰다. 감성 정보와 장르를 활용하여 TV프로그램을 감성 분류하는 실험을 한 결과, 감성 분류에 따라 다른 성능을 나타내긴 하였지만 일반적으로 장르와 감성 정보가 함께 활용한 경우가 장르나 감성 정보를 단독으로 활용된 경우보다 높은 감성 분류 정확도를 보여주었다. TV프로그램에 대해 감성 분석하는 방법과 감성 정보의 활용 가치를 보여주었지만 다양한 감성분류에 대한 기준을 세웠으나, 실제 각 감성별로 그 특성이 제대로 반영되지 않아, 분류 기준대로 감성분류가 명확하게 되지 않았다. 따라서 도메인에 대해 다양한 감성분류를 하기 위해서는 도메인에 적합한 감성 분류 기준이 필요하다.

이철성 외 3명의 연구에서는 영화에 대한 다양한 감성 분포를 알아보기 위하여 한글 문서를 기반으로 기계학습 모델을 적용하여 7개의 감성으로 분류하고 그 결과를 영화평에 적용하여 영화 장르별 감성특성을 분석하였다[5]. '다항 나이브 베이즈(Multinomial Naive Bayes)' 모델을 이용하여 네이버 40자 영화평에 대해 영화 100편에 해당하는 영화평의 감성을 분류하고, 요인분석(Factor analysis)하였다. 그 결과 감성 간의 관계를 파악하고 영화의 평점에 미치는 영향을 파악할 수 있었다. 이를 통해 영화에서 다양한 감성분석이 평점과 같은 기준이 될 가능성을 보여주었지만 그 유용성을 검증하지는 못하였다.

이경원 외 1명은 온라인 커뮤니티에서 사람들이 주고받는 주관적 정보는 소비자의 기대, 평가와 관련되어 흥행에 중요한 영향을 미친다는 결과를 바탕으로 영화리뷰의 감성 어휘를 분석하여 감성 분포맵을 제작하고 영화를 분석하였다[6]. 하지만 51개라는 많지 않은 감성 어휘로 분석하여 다양한 감성을 포괄하기 부족하고 측정한 감성 분포에 대한 활용 방안을 제시하지 못했다.

대부분의 연구들은 영화에 대한 리뷰나 마이크로블로그의 감성 분석을 통해서 영화에 대한 평점, 선호도, 흥행 예측 또는, 영화와 감성분석 결과와의 상관관계를 알아보는 연구에 그쳤다.

감성분석을 기반으로 영화간의 유사도를 측정하기 위해서는 영화에 대해 사람들이 느끼는 감성에 대해 도메인에 적합한 감성 분류 기준이 필요할 뿐만 아니라, 다양한 유사도 측정방식을 통해 가장 적합한 유사도 측정방식을 찾아내는 것이 필요하다.

3. 제안방법

본 절에서는 Thayer 모델을 기반으로 한 감성분류기준을 바탕으로 WordNet[2]을 이용하여 감성사전을 구축하고, 이를 이용하여 영화와 관련된 트윗의 감성분석을 통해 영화간에 감성에 대한 유사도를 측정하는 방법에 대해서 기술한다.

2.1 감성분석

감성분석을 수행하기 위해 어떤 상황에서 사람이 느끼는 감성을 다양한 형용사를 사용하여 분류하고자 많은 연구가 진행되고 있다. 일반적으로 감성을 차원적 공간으로 나타내는 방법은 감성의 가장 기본적인 특성을 단순화시켜 설명할 수 있기 때문에, 감성의 차이를 나타내고 구분하는데 있어 효율적이다.

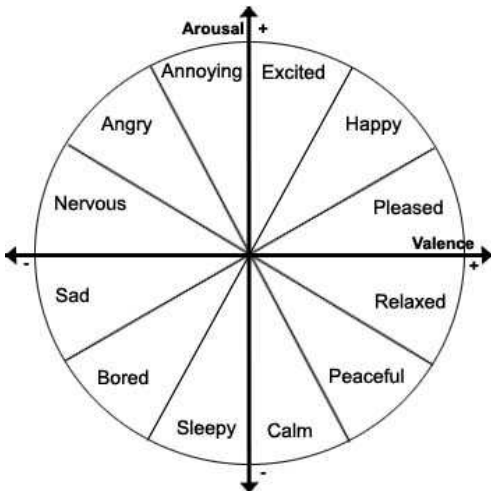

그림 2. Thayer 감성 모델
Fig. 2. Thayer's mood model

Thayer모델은 생물심리학적 관점에서 인간의 감성을 긍정, 부정의 정도를 나타내는 Valence와 활성 정도를 나타내는 Arousal 축으로 이루어진 2차원 평면에 대입함으로써 인간의 감성을 보다 직관적으로 서술하기 위한 대표적인 측정 모델로 주로 사용된다[7]. 따라서 영화에 대해 사람들이 느끼는 감성을 세분화된 기준으로 분류하기 위해서 Thayer 모델에서 제시된 12가지 감성클래스를 활용하여 감성의 분류기준을 정의하였다. 감성 분류 기준은 그림 2와 같이 Excited, Happy, Pleased, Relaxed, Peaceful, Calm, Sleepy, Bored, Sad, Nervous, Angry, Annoying 총 12개의 감성 클래스로 이루어져 있다.

세분화된 12개의 감성 클래스로 감성 분석을 하기 위하여, 각 감성어휘 및 감성어휘의 동의어(Synset)를 이용하여 감성사전을 구축한다. 이 때, WordNet을 이용하여 각 감성어휘에 대한 동의어를 정의한다. WordNet은 영어 의미 어휘목록 관계정보를 담은 사전으로써, 명사, 동사, 형용사와 부사를 각각 동의어 집합으로 분류하여 개념을 표현하였으며, 동의어 집합은 의미와 어휘 관계의 내부 링크로 연결되어 있는 네트워크로 구성된다. 이는 각 동의어 집합에 대한 상위어(hypernym), 하위어(hyponym), 등위어(coordinate term), 전체어(holonym)등의 의미관계를 제공한다.

표 1. Thayer 모델을 이용하여 생성된 감성사전 일부
Table 1. A part of Sentiment dictionary

Sentiment	Sentiment Lexicons
Excited	affect, affright, alter, arouse..
Happy	blessed, blissful, bright...
Pleased	care, delight, enchant, endear...
Relaxed	act, affect, alter, behave...
Peaceful	amicable, calm, dovish...
Calm	aplomb, appease, assuage...
Sleepy	asleep, dead, dormant, dozy...
Bored	tired, uninterested, arid, blunt...
Sad	bad, bittersweet, depressing...
Nervous	excitable, tense, troubled...
Angry	aggravated, angered, black...
Annoying	antagonize, beset, chafe...

감성분류 기준이 되는 12가지의 단어를 기반으로 WordNet을 이용하여 456개의 감성어휘로 감성사전을 확장시켰으며, 표 1은 WordNet을 활용하여 확장한 감성사전의 일부이다. 생성된 감성사전을 이용하여, 각 영화에 대한 트윗과 감성사전과의 어휘 유사성(lexical similarity)을 비교함으로써, 트윗에서 감성 단어가 포함되는 여부를 각 영화별로 취합하여 각 영화에 대한 세분화된 감성의 노출 빈도를 측정하였다.

식 (1)과 같이 트윗마다 노출되는 전체 감성 단어의 수에서 특정 영역의 감성 단어의 수의 비율을 계산하여 12개 감성의 비율을 계산하여 영화에 대한 감성벡터 $E_m = (e_1^m, \ldots, e_{12}^m)$을 구한다. E_m은 영화 m에 대한 감성 벡터를 나타내며 e_i^m는 영화 m에 대한 i번째 감성의 비율을 나타낸다. 이를 구하는 식은 식(1)과 같다. W_i는 i번째 감성에 대한 감성사전에 있는 어휘집합을 나타낸다.

2) http://wordnet.princeton.edu/

$T_m(w)$는 영화 m 과 관련된 트윗들에서 단어 w가 나타난 횟수를 의미한다.

$$e_i^m = \frac{\sum_{w \in W_i} T_m(w)}{\sum_{k=1}^{12} \sum_{w \in W_k} T_m(w)} \quad (1)$$

2.3 유사도 측정

영화의 장르에 대한 정보와, 영화와 관련된 트윗의 감성 분석을 통해 얻어진 영화에 대한 감성정보를 이용하여 영화간의 유사도를 계산한다. 장르와, 감성에 대한 유사도를 선형 결합을 통하여 최종 영화간의 유사도를 구한다.

$$Similarity(X,Y) = \alpha E(X,Y) + (1-\alpha) G(X,Y) \quad (2)$$

where

$$E(X,Y) = S(E_X, E_Y)$$
$$G(X,Y) = S(G_X, G_Y)$$

식 (2)에서 $Similarity(X,Y)$는 두 영화 X, Y의 유사도를 나타내고, $E(X,Y)$는 두 영화 X, Y에 대한 감성벡터 E_X와 E_Y 의 유사도를 나타낸다. $G(X,Y)$는 IMDB에서 제공하는 영화의 장르정보를 이용한 장르벡터 G_X와 G_Y에 대한 유사도를 나타낸다. 장르벡터는 총 27가지의 장르에 대하여 어떤 영화가 각 장르에 속한 여부를 0과 1로 표현한 것이다. α는 가중치로 0에서 1사이의 값을 가진다.

함수 S는 두 벡터 사이의 유사도를 측정하는 함수로, 본 논문에서는 유클리드 거리(Euclidean Distance), 코사인 유사도(Cosine Similarity), 피어슨 상관계수(Pearson correlation coefficient), 자카드 계수(Jaccard's coefficient)를 이용하였다.

2.3.1 유클리드거리(Euclidean Distance)

유클리드 거리는 두 점이 떨어진 거리를 측정하는 방법으로 사이 거리가 멀수록 유사하지 않은 것으로 판단한다. 서로 다른 두 영화 X, Y에 대하여 비교 정보가 장르벡터일 때 P는 27이며, 감성 벡터일 때 P는 12가 된다. 영화간의 유사도는 유클리드 거리가 클수록 유사도가 낮아져야 하므로, 유클리드 거리에 역수를 취한 값을 영화 X, Y에 대한 유사도로 정의한다. 유클리드 거리를 이용한 영화 X, Y의 유사도는 식 (3)과 같이 정의된다.

$$S_e(X,Y) = \left(\sum_{i=1}^{P} (x_i - y_i)^2 \right)^{-1/2} \quad (3)$$

2.3.2 코사인 유사도(Cosine Similarity)

코사인 유사도 계산은 내적공간의 두 벡터사이의 각도를 측정하여 유사한 정도를 구하는 방법이다. 식 (4)에서 서로 다른 두 영화의 비교 정보를 벡터 X, Y라 두고 두 벡터의 내적을 각 벡터 크기의 곱으로 나누어 $\cos\theta$를 구한다. 비교정보인 장르와 감성 비율이 모두 0과 1사이의 값이므로 코사인 유사도가 1이면 두 영화가 매우 유사하며, 유사도가 0이면 유사하지 않음을 나타낸다.

$$S_c(X,Y) = \frac{X \cdot Y}{\| X \| \, \| Y \|} \quad (4)$$

2.3.3 피어슨 상관계수(Pearson correlation coefficient)

두 변수 간에 어떠한 선형적 관계를 가지고 있는지 분석하는 방법으로 계산 결과가 1일 경우 강한 양적 선형 관계, 0일 경우 무시될 수 있는 선형 관계, -1일 경우 강한 음적 선형 관계를 의미한다. 두 영화 X, Y에 대해 식 (5)를 계산하여 각 영화의 비교 정보 간에 서로 상관이 있는지 분석한다. 비교 정보의 공분산을 표준편차로 나눠 선형 관계를 구한다.

$$S_p(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (5)$$

2.3.4 자카드 계수(Jaccard's coefficient)

자카드 계수는 두 데이터 집합의 교집합의 크기를 합집합의 크기로 나눈 것으로 장르의 경우, 대부분의 영화는 두 가지 이상의 장르에 속할 수 있으며 일반적으로 전체 장르 27개 중 해당 영화가 속하는 장르보다 속하지 않는 장르의 수가 많다. 따라서 비교할 두 영화 모두가 속하지 않는 장르를 제외하고 장르 간의 유사도를 측정한다. 두 영화 X, Y의 장르 집합을 각각 X, Y라 하고 식 (6)을 이용하여 장르에 대한 유사도를 구한다.

$$S_j(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (6)$$

3. 실험결과 및 분석

3.1 실험데이터

트윗을 수집할 영화 선정을 위해 IMDB 사용자 평점[3]을 기준으로 한 최근 인기 영화 목록에서 상위 400개 영화를 이용하였다. 데이터 수집 시 "Up", "Traffic" 등과 같이 제목이 한 단어로 구성되어 있는 제목의 영화인 경우, 영화와 관련 없는 트윗이 수집될 수 있으므로, 정확한 실험을 위해 400개 영화 중 제목이 두 단어 이상으로 이루어진 영화를 대상으로 하였다.

2013년 8월 10일부터 8월 30일까지 3주 동안 350개 영화에 대해서 860,154건의 트윗을 수집하였다. 또한 수집된 트윗에서 감성어휘가 100회 미만으로 나온 영화들은 감성유사도를 측정하는데 적합하지 않기 때문에 제외하였다.

그 결과 245개의 영화를 선정하였으며, 856,608건의 트윗을 실험데이터로 사용하였다. 평가지표로는 IMDB에서 제공하는 유사한 영화에 대한 정보를 이용하였다. IMDB에서는 특정 영화에 대해 유사한 영화를 12편에 대한 정보를 제공하고 있으며, 그 중 245개의 선정된 영화 내에서 IMDB에서 제공하는 유사한 영화와 일치하는 영화에 대한 통계자료는 표 2와 같다. 예를 들어 IMDB에서 제공하는 유사한 영화에 대한 정보가 245개의 영화 내에서 1편이 존재하는

3) http://www.imdb.com/list/9QxFVrhRpLI/

영화의 수가 총 58개임을 보여주고 있다.

표 2. IMDB에서 제공하는 유사 영화 통계
Table 2. The statistic of similar movies provided by the IMDB

The number of similar movies	The number of movies
0	67
1	58
2	54
3	39
4	21
5	6

3.2 실험결과

본 논문에서는 영화의 장르 정보와 영화와 관련된 트윗을 이용한 감성정보를 이용하여 영화간의 유사도를 측정한다. 영화의 장르는, IMDB에서 제공하는 영화에 대한 장르 정보를 이용하였다. 유클리드거리, 코사인유사도, 피어슨 상관계수, 자카드계수를 이용하여 영화 장르 정보 기반 영화간의 유사도를 구한다음, IMDB 결과와 비교를 통해 장르에 대해 가장 좋은 유사도 측정 방식을 알아내고, 이를 기반으로 감성에 대한 유사도를 계산하여 최종 영화간의 유사도를 구한다. 그리고 유사도 측정 방식 마다 측정된 유사도가 높은 영화 상위 5개를 IMDB에서 제공하는 유사한 영화 결과와 비교하여 성능을 평가하였다.

본 논문에서는 영화의 장르 정보만을 이용하여 영화간의 유사도를 구하는 방식을 베이스라인으로 두었으며, 표 3은 영화의 장르 정보만 이용하여 유사도 측정방법에 따라 추천된 영화 상위 5개가, IMDB의 유사 영화리스트 포함되어있는지 여부에 대한 결과를 보여주고 있다.

표 3. IMDB 유사 영화와 장르 기반의 유사 영화 결과 교
Table 3. The comparison of similar movies result based on genre of movies with the IMDB

IMDB result		Similarity Measures			
#Similar moives	#Movies	Euclidean	Cosine	Pearson	Jaccard
1	58	21	25	24	25
2	54	34	34	34	34
3	39	31	32	32	32
4	21	11	11	11	11
5	6	5	4	5	5
Total	**178**	**102**	**106**	**106**	**107**

장르 정보를 이용하여 영화간의 유사도를 측정한 결과, 자카드 계수를 이용한 결과가 가장 IMDB에서 보여준 결과에 가까운 결과를 보였다. 따라서 자카드 계수의 방법으로 구한 장르 유사도를 기반으로 영화에 관련된 트윗을 이용하여 감성분석하고, 유클리드 거리, 코사인 유사도, 피어슨 상관계수의 방법으로 구한 감성 유사도를 선형 결합을 통하여 최종 영화간의 유사도를 식 (2)와 같이 계산한다.

자카드 계수의 경우, 단순 매칭 계수를 이용한 것이기 때문에 트윗의 감성 분포와 같이 실수를 속성값으로 갖는 경우에 적합하지 않다. 따라서 감성 유사도를 측정하는 방식으로 자카드 계수는 고려하지 않았다. 가중치 α값 0.7로 설정하였으며, 앞선 실험과 마찬가지로 영화의 장르와 감성 이용하여 유사도 측정방법에 따라 추천된 영화 상위 5개가, IMDB의 유사 영화리스트 포함되어있는지 여부에 대한 결과를 보여주고 있다.

표 4. MDB 유사 영화와 장르와 감성 기반의 유사 영화 결과 비교
Table 4. The comparison of similar movies result based on genre and sentiment of movies with the IMDB

IMDB result		Similarity Measures		
#Similar moives	#Movies	Euclidean	Cosine	Pearson
1	58	25	30	28
2	54	29	31	37
3	39	28	31	33
4	21	11	11	13
5	6	5	5	5
Total	**178**	**98**	**108**	**116**

유사도 측정 방식에서 맞춘 영화 수를 IMDB의 유사 영화 수로 나눈 값을 정확도라고 하였을 때, 그림 2와 같이 장르에 대해서는 자카드 계수를 이용하여 유사도를 구하고, 감성에 대해서는 피어슨 상관계수를 이용하여 유사도를 구한 결과가 가장 좋은 성능을 보였다. 유클리드 거리의 경우, 단순히 장르 정보만을 이용한 결과보다 낮은 성능을 보였는데, 이는 감성의 유사도를 비교하는데 있어, 물리적인 거리를 측정하는 유클리드거리가 적합하지 않기 때문에 나타난 결과로 보인다. IMDB에서 제공하는 유사영화 정보와 하나 이상 일치하는 영화 수는 장르 정보만을 이용했을 때보다 본 논문에서 제안한 방식으로 했을 때, 8.4% 향상된 결과를 보였다.

표 5 . 장르 정보만을 이용한 방식과 장르와 감성 정보를 이용한 유사 영화 종합 결과 비교
Table 5. The comparison of total similar movies result based on genre of movies with based on genre and sentiment of movies

IMDB result		Similarity Measures	
#Similar moives	#Movies	Genre(Jaccard)	Genre(Jaccard) + Sentiment(Pearson)
1	58	25	28
2	108	50	51
3	117	50	54
4	84	30	32
5	30	18	13
Total	**397**	**173**	**178**

IMDB 유사 영화 리스트와 비교했을 때, 영화의 장르 정보만을 이용하여 유사도를 측정한 결과와 장르 정보 및 영화에 대한 감성 정보를 이용하여 유사도를 측정하여 추천된 상위 영화 5개 중 몇 개가 IMDB 유사 영화 리스트에 포함되어 있는지에 대한 결과는 표 5와 같다. 유사 영화 전체 수를 고려했을 때 약 2.9% 향상된 결과를 보였다.

연구에서 사용된 영화 장르 정보는 IMDB에서 제공하는

정보를 사용하였으며, IMDB에서 제공하는 유사 영화 정보는, IMDB에서 제공하는 장르 정보와 매우 큰 상관관계를 가지고 있음에도 불구하고, IMDB에서 제공하는 장르 정보만 이용한 결과보다 더 좋은 성능을 보였다. 이는 제안하는 방식은 장르에 국한되지 않고 감성 정보를 이용하여 다양한 장르의 영화를 추천한다고 보여진다.

4. 결론 및 향후 연구

본 논문에서는 영화간의 유사도를 비교하기 위하여 인기 영화 상위 400개에 대하여 영화에 대한 트윗을 수집하고, 이를 이용하여 세분화된 감성분석을 통해 영화간의 유사도를 구하는 방식을 제안했다. 세분화된 감성분석을 위해 Thayer 모델을 이용하여 12가지의 감성 분류 기준을 세우고 WordNet을 이용하여 감성사전을 확장하였다. 실험 결과 IMDB에서 제공하는 장르 정보에 대해 자카드 계수 방식을 이용하고, 트위터 감성분석을 통해 얻은 감성 정보에 대해 피어슨 상관계수를 이용하여 영화간의 유사도를 측정한 방식이 가장 좋은 성능을 보였다. 추후 감성 정보를 이용하여 추천된 유사 영화와, IMDB에서 제공하는 유사 영화 리스트에 대해 대중이 어떤 결과를 더 만족할지에 대한 판단이 필요하다. 영화에 대해 느낀 감성을 다양한 기준으로 분류함으로써 영화에 대한 세분화된 감성분석을 통해 추후, 영화 주연배우, 줄거리 등과 같은 영화에 대한 메타데이터와 함께 사용자에게 영화를 추천하는 추천시스템 등 다양한 분야에서 활용될 수 있을 것이다.

References

[1] Bing Liu, *Web data mining: exploring hyperlinks, contents, and usage data.* Springer, 2007.

[2] Grimes, Seth. "Sentiment Analysis: Opportunities and Challenges," Beye Network, 2008.

[3] Oghina, Andrei, et al. "Predicting imdb movie ratings using social media," *Advances in Information Retrieval,* Springer Berlin Heidelberg, pp. 503-507. 2012.

[4] Eggink, Jana, and Denise Bland. "A large scale experiment for mood-based classification of tv programmes," *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, IEEE, 2012.

[5] Cheolseong Lee, et al. "Classification and Analysis of Emotion in Korean Microblog Texts," *Journal of KIISE: database* vol.40, no.3, pp. 159-167, 2013.

[6] Lee, Kyung-won, and Ha, Hyo-Ji, "The Analysis on the Emotion Word and the Situation of Watching Movie to Develop Movie Recommender System," *Korean Society of Design Science Spring Symposium,* pp. 78-79, 2013.

[7] Thayer, Robert E. *The biopsychology of mood and arousal.* Oxford University Press, 1989.

저 자 소 개

김경민(Kyoungmin Kim)

2008 ~ 2013년 : 을지대학교 의료전산학 공학사

2013년 ~ 현재 : 성균관대학교 대학원 전자전기컴퓨터공학과 석사과정

관심분야 : Machine Learning, Sentiment Analysis

Phone : +82-31-290-7987

E-mail : kmkim1222@skku.edu

김동윤(Dong-Yun Kim)

2012년 ~ 현재 : 성균관대학교 컴퓨터공학과

관심분야 : Data Mining

Phone : +82-10-4120-5801

E-mail : picra0@gmail.com

이지형(Jee-Hyong Lee)

1993년 : 한국과학기술원 전산학과 학사

1995년 : 한국과학기술원 전산학과 석사

1999년 : 한국과학기술원 전산학과 박사

2002년 ~ 현재 : 성균관대학교 정보통신공학부 교수

관심분야 : Intelligence System, Machine Learning, User Modeling

Phone : +82-31-290-7154

E-mail : john@skku.edu