감정 분석을 이용한 협업적 영화 추천 방법

박한샘*, Abdel-Ilah Zakaria Khiati*, 강대현*, 권경락*, 정인정*
*고려대학교 컴퓨터정보학과
e-mail : {park11232000, le_zakkaz, internetkbs, helpnara, chung}@korea.ac.kr

Collaborative Movie Recommendation Method Using Sentiment Analysis

Hansaem Park*, Abdel-Ilah Zakaria Khiati*, Daehyun Kang*, Kyunglag Kwon*, In-Jeong Chung*
*Dept. of Computer and Information Science, Korea University

요 약

웹 2.0 의 폭발적인 성장과 스마트기기의 대중화 및 모바일 서비스의 활성화로 인하여 다양하고 방대한 양의 멀티미디어 콘텐츠가 보편화되었다. 따라서, 최근에 이를 효과적으로 활용하기 위한 다양한 연구가 수행되고 있다. 그러나, 사용자들은 아직도 수많은 멀티미디어 콘텐츠들 중에서 자신들이 원하는 콘텐츠를 찾는데 많은 어려움을 겪고 있다. 이에 따라, 사용자들의 올바른 의사결정을 도와주는 추천시스템에 대한 중요도가 나날이 급증하고 있다. 본 논문에서는 영화에 대해 사용자들이 남긴 리뷰로부터 감정 분석을 하고 분석된 각 사용자들의 감정 수치를 기반으로 영화추천 방법을 제안한다. 제안한 방법은 사용자들의 리뷰를 수집하고 각 사용자들의 감정 단어를 추출한다. 추출한 감정 단어들은 센티워드넷을 이용하여 사용자의 감정이 나타내는 정도를 분석한다. 분석된 사용자들의 감정 정보들을 바탕으로 사용자들에게 적절한 영화를 추천한다.

1. 서론

최근 스마트기기의 대중화와 웹 2.0 의 급속한 발전과 함께 시공의 제약 없이 온라인 서비스를 실시간으로 이용 할 수 있게 되었다. 이를 통해 많은 사용자들이 멀티미디어 콘텐츠를 공급하고 공유하는 것이 가능해 졌다. 그러나, 사용자들이 공급하는 수많은 멀티미디어 콘텐츠들로 인해 사용자는 자신에게 적합한 멀티미디어 콘텐츠를 찾는데 어려움을 겪고 있다[1].

많은 사용자들은 자신들이 좋아하거나 이전에 상영했던 영화들 중 자신에게 적절한 영화를 찾기 위해 영화 정보를 제공해주는 TV 프로그램이나 포털 사이트의 지식 커뮤니티를 이용하고 있다. 그러나, 이러한 방식은 사용자의 개인적인 취향을 완전히 반영하기가 어렵고 추천의 질이 낮은 단점이 존재하기 때문에 적절한 영화추천이 이루어지기 힘든 경우가 많다[2].

이와 같은 문제를 해결하기 위해 Netflix[1], IMDb[2], Flickr[3] 등은 자동화 된 추천 시스템을 제공하고 있다. 이 추천 시스템들은 대부분 사용자들의 평가를 기반으로 평가 등급이 높을 경우에 다른 사용자에게 높은 평가를 받은 콘텐츠들을 추천하는 방식을 사용하고 있는데, 이러한 추천 시스템들은 데이터 희소성(Data Sparseness)문제를 가지고 있다[3]. 데이터 희소성 문제란 사용자들에 의해 평가되어지지 않은 콘텐츠가 있을 경우에 추천시스템에서 다른 사용자들에게 올바른 추천을 하지 못하는 문제를 말한다.

이러한 문제점을 극복하기 위해 최근 수행되고 있는 추천 시스템에 관한 연구들에서는 추천 시스템에 사용자의 리뷰나 코멘트들로부터 사용자의 의견이나 감정 등을 파악하는 감정 분석이 많이 활용 되고 있다[4, 5]. 또한, 감정 분석은 추천 시스템에서 사용자가 느끼는 주관적인 정보를 파악하는데 효과적이다. 따라서, 영화 추천 시스템에서 개인 맞춤형 영화를 추천하기 위해서는 사용자들의 리뷰를 가지고 감정 분석을 하고 이를 바탕으로 사용자들의 주관적인 정보를 분석해야 한다. 본 논문에서는 사용자들의 리뷰나 코멘트들로부터 감정 단어를 추출하고 센티워드넷[4]을 활용하여 추출된 감정 단어에 대한 감정 정도(Sentiment Degree)를 분석 한 후, 효과적으로 영화를 추천하는 방식을 제안한다. 마지막으로 제안한 방법을 이용하여 사용자들에게 개인 맞춤형 멀티미디어 콘텐츠를 추천할 수 있음을 파일럿 실험을 통하여 보인다.

[1] https://www.netflix.com
[2] http://www.imdb.com
[3] http://www.flickr.com
[4] http://sentiwordnet.isti.cnr.it

2. 관련 연구

2.1. 감정 분석(Sentiment Analysis)

오피니언 마이닝(Opinion Mining)이라고도 불리는 감정 분석에 대한 연구의 목적은 콘텐츠에 대한 사용자의 선호도를 표현하기 위해 어떤 단어들이 사용되는지를 분류하는 것이었다[4]. 이와 같이, 감정 분석에서 가장 중요한 일 중 하나는 바로 어떤 단어들이 사용자들의 감정을 표현하는지를 찾아 내는 것이다.

최근 들어, Tripadvisor[5], Expedia[6], Amazon[7]과 같은 전자상거래 사이트들에서도 사용자들에게 개인 맞춤형 추천을 제공하기 위해서 콘텐츠에 대한 사용자의 리뷰나 코멘트들을 수집하고 있으며 이를 위한 다양한 서비스들이 개발되고 있다[5]. 일반적으로, 감정 분석은 문서 안에 표현된 단어들을 사실과 의견 2 가지 종류로 분류한다. 사실이란 객체나, 실제 일어난 일에 대한 객관적인 표현을 말한다. 그리고 의견이란 이런 객체나 실제 일어난 일들에 대한 감정, 태도, 느낌과 같은 주관적인 표현들을 말한다. 또한, 어떤 객체나 실제 일어난 일에 대한 사용자의 주관적인 표현들은 사용자의 감정을 긍정(Positive), 중립(Neutral), 부정(Negative)으로 표현한다. 따라서, 사용자들의 감정을 분석하기 위해서는 사용자의 리뷰나 코멘트에서 감정 단어를 추출하는 것이 중요하다[6].

본 논문에서는 사용자의 리뷰나 코멘트로부터 사용자 감정에 영향을 주는 단어들을 수집하기 위해서 자연어 처리(Natural Language Processing) 과정을 수행하고 센티워드넷을 통해 감정 단어의 정도를 파악한다.

2.2. 추천 시스템

전자상거래나 집단지성 기반의 웹사이트에서의 콘텐츠 추천은 사용자가 원하는 물건을 찾고 구입하는데 있어서 적절한 의사결정을 하는데 중요한 역할을 하고 있다. 콘텐츠를 추천하는 방법은 크게 콘텐츠 기반 추천 방법(Content-Based Recommendation), 협업적 추천 방법(Collaborative Filtering), 이들을 혼합한 혼합형 추천 방법(Hybrid Recommendation)으로 나뉜다[7].

콘텐츠 기반 추천 방법은 사용자가 과거에 좋아했던 콘텐츠와 유사한 콘텐츠를 추천해 주는 방법이다. 반면 협업적 추천 방법은 과거에 같은 선호도나 취향을 가지고 있는 사용자들이 좋아했던 콘텐츠들을 추천해주는 방법을 말한다. 또한, 협업적 추천 방법은 다시 아이템 기반 추천 방법[8]과 사용자 기반 추천 방법[9]으로 나눈다. 그러나 위와 같은 추천 방법들은 평가되지 않은 콘텐츠나 새로운 사용자가 추가 되었을 때는 콘텐츠 추천의 질이 낮아지는 단점을 가지고 있다. 따라서, 본 논문에서는 수많은 사용자들의 감정을 분석한 협업적 추천 방법을 이용하여 위의 문제점들을 해결하고자 한다.

[5] http://www.tripadvisor.co.kr
[6] http://www.expedia.co.kr
[7] http://www.amazon.com

3. 제안하는 방법

본 논문에서 제안하는 방법은 다음과 같다. 대표적인 영화 포털 사이트인 Internet Movie Database (IMDb)를 통해 사용자의 정보, 사용자가 영화에 대해 작성한 리뷰들과 영화에 대한 정보들을 수집한 후, 사용자 리뷰로부터 사용자들의 감정을 나타내는 감정 단어를 추출한다. 그리고, 센티워드넷을 이용하여 추출된 감정어가 가지고 있는 감정 정도를 분석한다. 분석된 여러 사용자들의 영화에 대한 감정 정도 그리고 영화 정보와 사용자 정보를 바탕으로 다른 사용자에게 영화를 추천한다.

제안한 방법은 크게 1) 데이터 수집, 2) 자연어 처리를 이용한 감정 단어 추출, 3) 센티워드넷을 이용한 감정 단어 분석 및 영화 추천으로 크게 3 단계로 나눈다.

3.1. 데이터 수집

본 단계에서는 IMDb 로부터 사용자 정보, 영화 정보, 그리고 영화에 대한 사용자의 리뷰 등의 데이터를 수집한다.

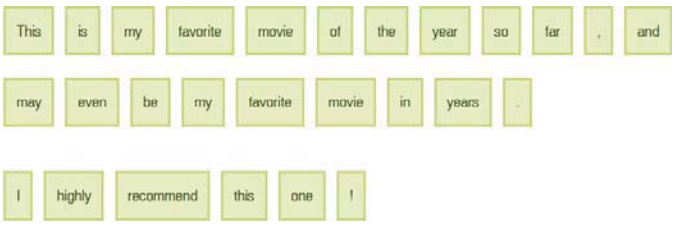

I can't wait to see it again!
Author: hesslop39 from Hollywood, CA
26 November 2013
The last time Disney adapted a Hans Christian Anderson fairy tale, we got one of my favorite films, The Little Mermaid. Now, we have a movie that has very much the same feel as Mermaid. Frozen feels like it could fit right in with the Disney movies of the late 80's/early 90's, like The Little Mermaid, Beauty and the Beast and The Lion King. The music is great, and I left singing it; the lead characters are very likable, with memorable sidekicks; and there are even surprises, things you would never expect to see in a Disney animated film! This is my favorite movie of the year so far, and may even be my favorite movie in years. I highly recommend this one!

(그림 1) IMDb 의 영화에 대한 사용자들의 리뷰

그림 1 과 같이 IMDb 에서는 영화에 대한 사용자들의 리뷰와 사용자의 정보 및 영화 정보를 제공한다.

3.2. 사용자 리뷰에서 감정 단어 추출

이전 단계로부터 수집된 사용자들의 리뷰로부터 사용자들의 감정과 연관된 단어들을 추출하기 위해 자연어 처리(Natural Language Processing) 도구인 NLTK (Natural Language Toolkit)[14]를 사용한다. NLTK 는 파이썬(Python)에서 자연어를 처리하기 위해 제공해주는 라이브러리를 말한다.



(그림 2) NLTK 를 이용한 사용자 리뷰에서 감정 단어 추출의 예

그림 2 와 같이 NLTK 를 이용하면 사용자의 리뷰로부터 감정 단어를 추출할 수 있다. 감정 단어가 될 수 있는 단어들은 명사, 동사, 형용사가 있다. 그러나 명사와 동사보다는 형용사가 사용자의 감정을 명확히

나타낸다. 따라서, 본 단계에서는 특정에 영화에 대한 여러 사용자가 남긴 리뷰로부터 형용사를 중심적으로 추출하고 파악한다.

3.3. 센티워드넷(SentiWordNet)을 이용한 감정 단어 분석 및 영화 추천

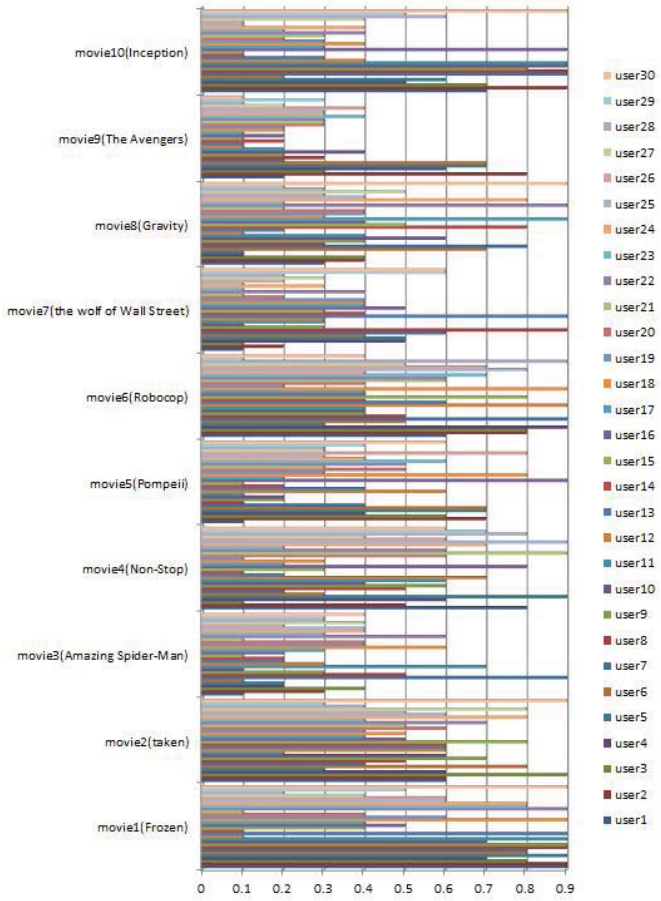
센티워드넷은 워드넷(WordNet)과 반지도 학습(Semi-Supervised Learning)을 이용한 어휘 사전이다. 따라서, 본 단계에서는 이전 단계로부터 추출된 감정 단어들의 극성(Polarity)을 분석한다. 감정 단어의 극성은 긍정(Positive), 중립(Neutral), 부정(Negative)으로 나타낸다. 그림 3 은 전 단계에서 추출된 감정 단어 'favorite'의 감정 정도를 나타낸 것이다. 이 단어는 긍정(0.125), 중립(0.875), 부정(0)의 정도를 가지고 있다.

그러나, 본 논문에서는 단어가 가지고 있는 중립과 부정의 감정 정도는 이용하지 않고 긍정의 감정 정도만을 이용한다. 예를 들어, 긍정의 감정 값이 0.4 이하일 경우에는 감정 정도가 낮다고 판단하고 0.6 이상일 경우에는 감정 정도가 높다고 판단한다. 따라서, N 명의 사용자들이 특정 영화에 대한 감정 정도가 0.6 이상인 영화만을 분석해서 추천한다.

(그림 3) 센티워드넷을 이용한 감정 단어의 정도를 분석한 예

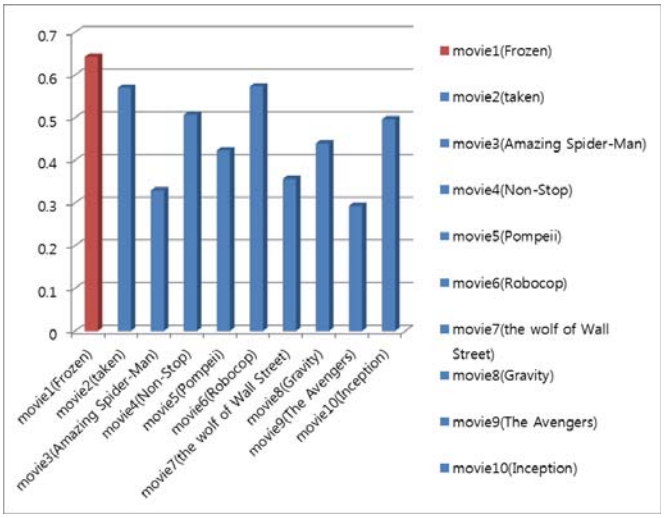
4. 파일럿 실험 및 결과

본 논문에서는 파일럿 실험을 위한 데이터로 IMDb 웹사이트로부터 임의의 사용자 30 명이 10 개의 영화에 대해 작성한 리뷰를 수집한다. 제안한 방법을 이용하여 이에 대한 감정 분석을 수행한다. 그림 5 는 임의의 사용자 30 명이 사용자가 10 개의 영화에 대한 감정 정도를 표현한 그래프이다.



(그림 5) 특정 영화에 대한 사용자의 감정 정도

그림 6 은 10 개의 영화에 대한 사용자의 감정 정도의 평균값을 나타낸다.



(그림 6) 영화에 대한 사용자의 감정 정도의 평균값

그림 6 에서 나타난 바와 같이 10 개의 영화에 대한 사용자의 감정 정도 평균값이 제일 큰 Frozen 이라는 영화를 다른 사용자에게 추천한다.

5. 결론 및 향후 과제

본 논문에서는 사용자의 영화에 대한 감정 정도를 분석하는 협업적 영화 추천 방법을 제안하였다. 파일럿 실험에서는 제안한 방법을 이용하여 특정 영화에 대한 여러 사용자의 감정을 분석한 후, 다른 사용자에게 적절한 영화 추천이 가능함을 보였다.

향후 연구로는 다양하고 더 많은 양의 데이터를 수집하여 제안한 방법의 성능을 평가하고 타당성을 검증할 것이다. 또한, 온라인 분석처리(OLAP)의 슬라이스(Slice), 다이스(Dice), 롤업(Roll-up), 드릴 다운/업(Drill-down/up)과 같은 다른 기능을 이용하여 사용자와 영화의 패턴을 다차원적으로 분석함으로써 새로운 사용자에게 효과적으로 영화를 추천하는 시스템을 구축하고자 한다.

사사

본 논문은 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 BK21 플러스 사업의 연구 결과임 (No. T1300571)

참고문헌

[1] P. C. George Lekakos, "A hybrid approach for movie recommendation," *Multimedia Tools and Applications,* vol. 36, pp. 55-70, 2008.

[2] 김부성, 김희라, 이재동, and 이지형, "사용자 개인정보를 이용한 협업 필터링 기반 영화 추천 시스템," *한국지능시스템학회 학술발표 논문집,* vol. 23, pp. 63-64, 2013.

[3] G. Guo, "Integrating Trust and similarity to Ameliorate the Data Sparsity and Cold Start for Recommender Systems," *RecSys '13 Proceedings of the 7th ACM conference on Recommender Systems,* pp. 451-454, 2013.

[4] C. W. Leung, S. C. Chan, and F.-l. Chung, "Integrating collaborative filtering and sentiment analysis: A rating inference approach," in *Proceedings of The ECAI 2006 Workshop on Recommender Systems*, 2006, pp. 62-66.

[5] C.-C. Musat, Y. Liang, and B. Faltings, "Recommendation using textual opinions," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 2684-2690.

[6] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," 2013.

[7] H. Ji, J. Li, C. Ren, and M. He, "Hybrid collaborative filtering model for improved recommendation," in *Service Operations and Logistics, and Informatics (SOLI), 2013 IEEE International Conference on*, 2013, pp. 142-145.

[8] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," *ACM Transactions on Information Systems (TOIS),* vol. 22, pp. 143-177, Jan 2004.

[9] P. Wang and H. Ye, "A personalized recommendation algorithm combining slope one scheme and user based collaborative filtering," in *Industrial and Information Systems, 2009. IIS'09. International Conference on*, 2009, pp. 152-154.

[10] A. Berson and S. J. Smith, *Data warehousing, data mining, and OLAP*: McGraw-Hill, Inc., 1997.

[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques: Concepts and Techniques*: Elsevier, 2011.

[12] A. Cuzzocrea, L. Bellatreche, and I.-Y. Song, "Data warehousing and OLAP over big data: current challenges and future research directions," in *Proceedings of the sixteenth international workshop on Data warehousing and OLAP*, 2013, pp. 67-70.

[13] A. Cuzzocrea, I.-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!," in *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, 2011, pp. 101-104.

[14] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*, 2006, pp. 69-72.