CogTV를 위한 생체신호기반 시청자 선호도 모델

A Viewer Preference Model Based on Physiological Feedback

박태서[1] · 김병희[2] · 장병탁[2]
Tae-Suh Park, Byoung-Hee Kim, and Byoung-Tak Zhang

[1]서울대학교 인지과학협동과정, [2] 컴퓨터공학부
E-mail: {taesuh, bhkim, btzhang}@bi.snu.ac.kr

요 약

본 논문은 TV를 이용한 영화시청 환경에서 해당 컨텐트에 대한 시청자의 암묵적 반응과 컨텐트의 멀티모달 피쳐를 실시간으로 측정 및 동기화하여 이를 기반으로 동영상 선호모델을 지속적으로 개선하고 필요시 영화추천을 수행하는 시스템을 제안한다. 제안한 컨텐트-시청자 연계 추천모델의 일례로서 컨텐트의 오디오 및 자막 정보로부터 시청자의 피부전기활성도로 측정된 arousal반응을 예측할 수 있음을 보인다.

키워드 : Recommendation System, Sentiment Analysis, EDA, Machine Learning, Physiological feedback.

1. 서 론

컨텐츠에 대한 사용자의 선호를 추정하는 기술은 collaborative filtering의 개발에 힘입어 도서, 영화 등 다양한 컨텐츠의 추천을 가능하게 했으나, 신규 사용자 혹은 신작영화처럼 참조할 학습데이터가 존재하지 않거나, 사용자의 선호도를 측정하기 어려운 환경에선 적용이 불가능하다는 문제점이 알려져 있다[1]. 특히, TV시청환경에서는 “lean-back”으로 대표되는 사용경험 특성상 시청자의 명시적인 만족도 피드백을 기대하기 어렵고, 프라이버시에 민감하여 시청자 개인정보나 시청만족도 자체를 수집하기 어렵다는 제약이 존재한다. 이러한 제약을 극복할 수 있는 방법으로 컨텐츠 내용에 기반한 추천모델이 연구되어 왔으나, 보편적인 선호요인을 찾기가 어렵고 profiling의 어려움으로 인해 collaborative filtering 등장 이후 보완적인 수단으로만 쓰여왔다[1].

본 논문에서는 상술한 문제를 극복하기 위해 TV시청환경에서의 시청자가 별도의 의지적 행동 없이 시청 중의 정서적 변화에 따라 무의식적으로 내비치는 암묵적 피드백(implicit feedback)을 추천모델 개선 및 대안컨텐츠 추천시점 결정에 활용하는 방법을 제안한다. 암묵적 피드백은 시청 중에 발생하는 시청자의 동영상에 대한 비언어적이고 수동적인 반응으로서, 시선과 자세, 생리적 반응을 포괄한다. 암묵적 피드백은 통상 별점으로 대표되는 명시적 피드백과 달리 측정 자체에 불확실성이 따르지만, 상대적으로 설문방식에 따르는 응답자의 의식적 왜곡을 피할 수 있다는 장점이 있다. 예를 들어 특정 시청자가 강한 흥미를 보이고 집중해서 시청한 영화의 구성요소를 분석하여 해당 시청자에게 특화된 선호모델(preference model)을 학습할 수 있고, 더 나아가 시청자의 집중여부를 지속적으로 모니터링하여 최적의 대안제시 시점을 결정할 수 있다.

감사의 글 : 본 연구는 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734), 산업통상자원부의 재원으로 한국산업기술평가관리원의 지원(KEIT-10035348)을 일부 받았음.

2. 암묵적 피드백 측정

Arousal과 Valence는 심리학 분야에서 인간의 다양한 감정을 포괄적으로 기술할 수 있도록 제안된 프레임웍의 두 축으로서, 본 논문에서는 Arousal측정에 초점을 맞춰 Arousal로 대표되는 시청자의 몰입도 및 감정적 고양수준을 해당 컨텐트에 대한 만족도의 중요한 지표로 간주하고, 이를 시청자의 피부에서 접촉식 센서에 기반한 피부전기활성(Electrodermal Activity, 이하 EDA)을 통하여 측정하고자 한다[2]. 그림1은 본 연구에서 사용된 시스템을 통해 취득한 영화시청 중 EDA수준과 자막감성수준의 시계열 데이터 중 일부이다.

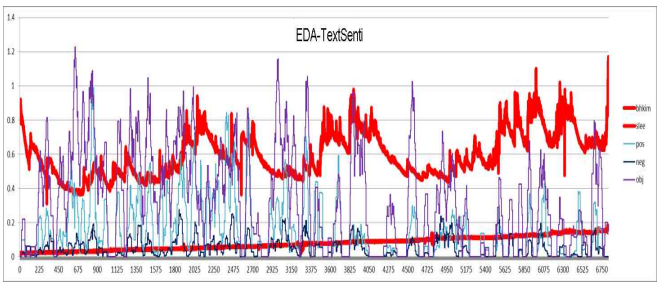

그림 1. 동일영상 시청 중인 피험자 두 명의 EDA반응(굵은선) 및 영상 자막 기반 세 가지 감성 지수(가는선)

또다른 만족도 지표로 시청자의 시선집중 수준을 사용할수있다. 구체적으로 특정 컨텐트의 시청중 임의 시구간 내에서의 시선(eye gaze)이 시청면을 향하는 시간의 점유율 및 깜박임 정도를 측정함으로써 해당 컨텐트에 대한 주의집중 수준을 정량적으로 측정할 수 있다.

3. 시스템 구성

제안하는 시스템은 TV를 이용한 동영상 시청 환경에서, 동영상 컨텐츠를 표현하는 다양한 피쳐 및 시청자의 암묵적 반응을 실시간에 측정 및 동기화하여 이를 기반

으로 시청자의 동영상 선호모델을 지속적으로 개선하고 필요시 영화추천을 수행하는 것을 목표로 한다.

시청자 암묵적 피드백 Y는 만족/선호모델 상태(R)의 각 생체신호별 특성(g)에 따른 발현으로서 콘텐트 구성요소 X와의 관계는 다음과 같이 기술된다:

$$\vec{R} = f(\vec{X};\theta)$$
$$\vec{Y} = g(\vec{R}+e) = g(f(\vec{X};\theta)+e) \quad (1)$$

여기서 g는 개인차가 상당한 발현수준 및 소요시간의 함수이나, 개인별 추정으로 한정하고 샘플링 시구간을 충분히 크게 잡으면 모델 파라미터의 학습과정은 아래와 같이 근사된다.

$$\hat{\theta} = \arg\min \sum_i |r_i - \tilde{r_i}| \approx \arg\min \sum_i |y_i - \tilde{y_i}|$$
$$where\ r \in \vec{R}, y \in \vec{Y} \quad (2)$$

따라서, 모델 파라미터가 주어진 상태에서의 추천은 백그라운드에서 취득된 대안 컨텐츠들의 R 혹은 Y의 추정값을 토대로 상위 N개를 후보로 제안하게 된다. 상술한 컨텐츠 구성요소 X에 따른 만족/선호모델 온라인학습을 지원하는 추천시스템의 구조는 그림 2과 같다. 암묵적반응 측정에는 3D 카메라, 아이트랙커 및 접촉식 EDA센서를 기본으로 사용하되 용도에 따라 확장가능하고(그림 2,3), 동영상에서의 다양한 요인추출 및 시청자 반응과의 연관성분석을 위한 컴퓨팅 시스템이 포함되어 있다.

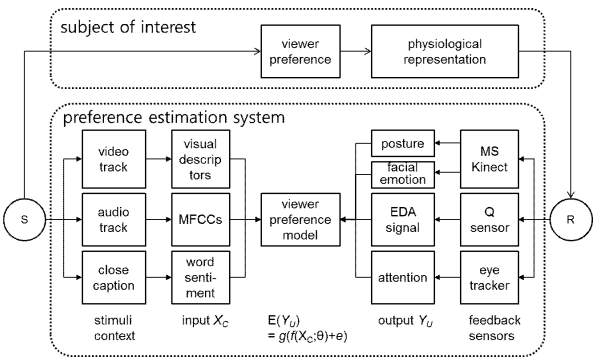

그림 2 시스템 구성도

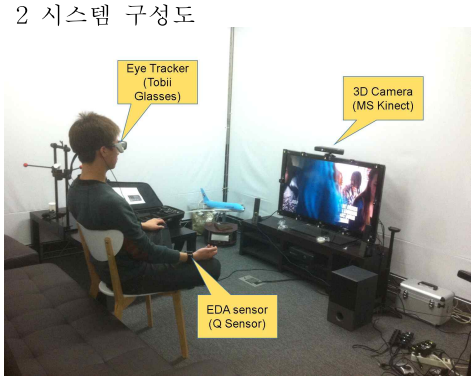

그림 3. 시청자의 암묵적 반응 측정을 위한 시스템 구축사례

4. 실험설계 및 결과

본 논문에서는 오디오-텍스트 피쳐를 기반으로 시청자 arousal수준을 추정한 실험사례를 일례로서 소개한다.

4.1 실험설계

20~30대의 여성 2명과 남성6명에게 각각 1편의 임의선정된 2시간 분량의 영화를 보여주고, 시청 중 매 25ms마다 취득한 EDA 데이터와 동영상에서 추출한 피쳐를 동기화시킨 학습-평가 데이터셋을 구축하였다. 동영상의 피쳐 중 오디오 정보인 MFCCs (Mel-Frequency Cepstral Coefficients)는 25ms 프레임 별로 20단계 필터 계수 및 각 필터에 대한 차분값(delta)을 추출하였으며, 롱텀 피쳐로서 5초 길이, 50% 중복 속성의 이동 윈도우를 적용한 후 윈도우별로 필터별 피처군의 1~4차 모멘트를 추출하였고, 텍스트 정보는 영화 자막 혹은 TV방송신호의 close caption트랙에서 추출된 문자열을 Python Natural Language Toolkit[3]기반 negation등의 전처리를 수행한 후 이를 SentiWordNet[4] 조회를 통하여 각 관심 시구간 내 감성요소(sentiment)를 세 가지 준위(pos, neg, obj)로 변환하였고, (pos-neg) 및 sqrt(pos^2 +neg^2) 피쳐 2종을 추가하였다. 피험자 반응은 착용형 EDA센서(Q Sensor)를 통해 8Hz로 샘플링된 시계열값을 상기 각 모달리티별 윈도우 내의 평균값으로 변환하여 동기화시켰다.

4.2 실험결과

그림 4는 8명의 피험자 대상 측정실험 각각에 대하여 오디오/텍스트 피쳐를 기반으로 EDA를 예측하는 모델을 기계학습의 감독학습 기법으로 구축한 결과의 비교 그래프이다. 두 경우 공히 개인별 EDA반응의 1사분위수 및 3사분위수를 기준으로 구분한 binary EDA class (high, low)를 Random Forest기법으로 학습하였으며, 평가시 3-fold cross validation을 5회 적용하여 통계적 비교를 수행하였다. 관찰된 Box Plot 패턴은 오디오정보가 텍스트 감성요소 정보대비 적어도 Accuracy관점에서 기계학습에 보다 적합하고 개인차도 상대적으로 적은 경향성을 보인다. 오디오정보를 이용할 경우, 피험자에 따라 약 73~88%의 Accuracy로 개별EDA를 예측할 수 있었다.

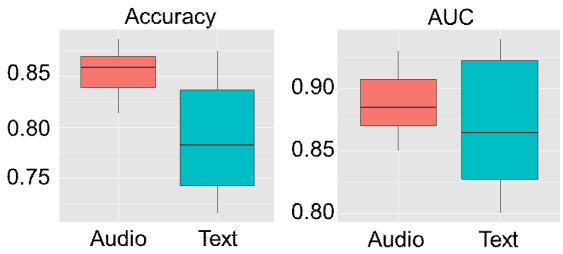

그림 4 오디오-텍스트기반 추정성능비교

실험결과, 자막기반 감성분석과 저수준 오디오분석 둘 다 시청자의 EDA예측에 활용가능하고, 상대적으로 저수준 오디오의 정보량이 더 많음을 확인할 수 있었다. 다만, 개인차가 큰 EDA특성상 예측값은 특정 개인에 대해서만 유효하다.

참고문헌

[1] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. Springer US, 2011.

[2] D. McDuff et al. "AffectAura: an intelligent system for emotional memory," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 2012.

[3] S. Bird, E. Loper and E. Klein, Natural Language Processing with Python. O'Reilly Media Inc., 2009.

[4] S. Baccianella, 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*, 2010.