다단계 분류기의 사용자 프로파일 추론을 통한 프로토타입 표적 광고 시스템 개발

김문조*, 이범식*, 임정연*, 김문철*, 이희경**, 이한규**
한국정보통신대학교 공학부 멀티미디어트랙*
한국전자통신연구원 디지털방송연구단 방송미디어연구그룹**

Development of Prototype Target Advertisement System Using a Viewer's Profile Reasoning with Multistage Classifier

Munjo Kim*, Bumshik Lee*, Jeongyeon Lim*, Munchurl Kim*, Heekyung Lee**, Hankyu Lee**
Multimedia Track Engineering Department
Information and Communications University*
Broadcasting Media Research Group in Digital Broadcasting Research Division
Electronics and Telecommunications Research Institute**
E-mail : *{kimmj, bslee, jylim, mkim}@icu.ac.kr, **{lhk95, hkl}@etri.re.kr

Abstract

In the uni-directional broadcasting environments, almost TV programs are scheduled depending on the viewers' popular watching time, and the advertisement contents in these TV programs are mainly arranged by the popularity and the ages of the audience. However, the advertisement programs which support the TV programs the audiences want are not served to the appropriate audiences efficiently. In this paper, we propose the prototype of target advertisement system for the appropriate distribution of the advertisement contents. The proposed target advertisement system estimates the audience's profile without private information and provides the target advertisement contents by using his/her inferred profile. And we show the accuracy of the proposed algorithm, Multistage Classifier, for the target advertisement system and the implementation of our target advertisement system.

I. 서론

현재 존재하고 있는 사용자 시청환경에서 대개의 광고는 방송사에 의해 프로그램의 인기도 혹은 시청률, 시청 연령, 시간대에 따라서 획일적으로 광고를 보내고 있다. 이런 무분별한 광고 제공 문제를 해소하기 위하여 많은 연구들이 진행되고 있지만, 기존의 연구들은 셋탑 박스(Set-top box)나 인터넷을 통하여 시청자가 프로파일(Profile)을 명시적으로 입력하고, 이렇게 입력된 정보를 바탕으로 광고 콘텐츠 제공자들이 표적 광고 서비스를 하고 있다[1-3]. 그러나, 인터넷을 통한 개인 정보 유출에 의하여 알지 못하는 사람으로부터 개인 정보가 악용 될 수 있기 때문에 시청자들은 개인 정보를 입력하는 것을 꺼려하는 것이 추세이다.

본 논문에서는 시청자의 개인 정보 유출 없이 시청자의 TV 시청 데이터(TV Usage History)와 다단계 분류기(MSC: Multistage Classifier) 알고리즘을 이용하여 시청자의 프로파일(성별, 연령대)를 추론하고 추론된 프로파일 결과를 바탕으로 표적광고를 제공한다.

II. 표적 광고 시스템 구조 (Target Advertisement System Architecture)

표적 광고 시스템은 방송국, 광고주, 시청자를 고려하며, 시청자의 TV 시청 데이터를 분석하여 시청자의 프로파일을 추론하는 추론기능, 추론된 결과를 바탕으로 표적 광고를 전송하기 위한 전송 기능, 사용자 인터페이스 기능이 필요하다. 그림 1 에서는 본 논문에서 제안하는 표적 광고 시스템의 구조를 보여주고 있다.

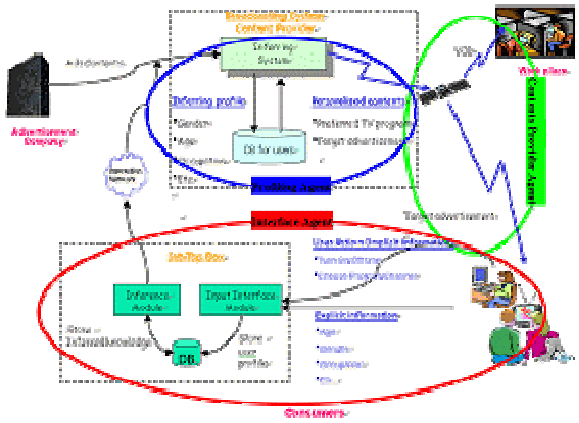

그림 1. 표적 광고 시스템 구조

그림 1 의 프로파일링 에이전트(Profiling Agent)는 사용자 인터페이스 에이전트로(Interface Agent)부터 저장된 TV 시청 데이터를 전달 받아 시청자의 성별 및 연령대를 추론하며, 추론된 시청자의 성별 및 연령대 정보를 기반으로 표적광고 콘텐츠를 선별한다. 선별된 광고 콘텐츠는 콘텐츠 공급자 에이전트(Contents Provider Agent)를 거쳐 사용자 인터페이스 에이전트를 통해 시청자가 콘텐츠 소비를 가능하게 한다.

III. 다단계 분류기(Multistage Classifier)를 이용한 사용자 프로파일 추론 알고리즘

3.1 특징 벡터(Feature Vector) 추출

본 논문에서 정의하고 있는 TV 시청 데이터는 전자 프로그램 가이드(이하, EPG)에서 제공하는 EPG 정보로부터 TV 시청 데이터를 가지고 있는 데이터베이스(이하, DB)를 의미한다. 일반적인 TV 시청 데이터의 데이터베이스에는 표 1 과 같은 필드(Field)들이 존재한다.

필드 이름	설 명
id	시청자 아이디
profile	시청자의 성별 및 연령
date	프로그램 방송 날짜
dayofweek	시청한 요일
subscstart_t	시청자 프로그램 시청 시작 시간
subscend_t	시청자 프로그램 시청 종료 시간
programstart_t	프로그램 본 시작 시간
programend_t	프로그램 본 종료 시간
title	프로그램의 제목
channel	프로그램의 채널(6 개 채널)
genre	프로그램의 장르(8 개 장르)

표 1. TV 시청 데이터 DB 의 필드 이름 및 설명

표 1 의 DB 로부터 서버에서는 시청자의 프로파일(성별 및 연령대)를 추론하기 위한 특징 값(Feature Value)은 다음 조건을 만족해야 한다.

- $\dfrac{(subscend\_t - subscstart\_t)}{(programend\_t - programstart\_t)} \geq T_{Th}$
- 특정 기간 동안 프로그램 시청 횟수$\geq C_{Th}$

위의 두 조건을 만족하면서 DB 에서 개인별로 추출할 수 있는 특징 값들과 개수는 표 2 와 같다.

특징 값 종류	개수
시청한 장르 횟수의 장르 확률(GPRC)	8
시청한 장르 시청시간의 장르 확률(GPRT)	8
특정 기간 동안의 평균 시청 시간(AVT)	1
시청한 채널 시청시간의 채널 확률(CPR)	6

표 2. 특징 값의 종류와 개수

표 2 의 특징 값들을 이용하여, 특정 요일에 해당하는 특징 벡터(Feature Vector)를 만들면 표 3 과 같다.

Index	1 ~ 8	9 ~ 16	17	18 ~ 23
특징 값	GPRC	GPRT	AVT	CPR

표 3. 특징 벡터(Feature Vector)

특징 벡터의 종류는 2 가지가 있는데, 시청자 별 특징 벡터와 성별과 연령대에 의한 그룹을 대표하기 위한 특징 벡터가 있다. 서버에 있는 프로파일링 에이전트는 시청자 별 특징 벡터를 가지고 있는 룩업 테이블(Look-Up Table)을 만들 수 있다. 다단계 분류기(이하, MSC)는 시청자에 따른 특징 벡터를 가지고 있는 룩업 테이블로 구성된 테이블을 이용하여 특정 시청자의 프로파일을 추론하게 된다. 또한, 특징 벡터는 주중의 데이터, 즉 월요일부터 금요일까지의 특징 벡터만 이용한다. 이는 주말에는 성별과 연령대에 따른 특징 값들이 유사하기 때문에 특징 값들로 이용할 수 없다.

3.2 제 1 단계 분류기(1st Stage Classifier)

제 1 단계 분류기는 특정 요일에 해당하는 특징 벡터간의 거리를 이용하여 구한다. 특징 벡터간의 거리 측정은 벡터 상관법(Vector Corrleation)과 정규화된 유클리드 거리(Normalized Euclidean Distance) 방법을 이용하여 특징 벡터간의 유사성을 측정한다. 벡터 상관법은 벡터간의 유사성 측정을 위하여 (식 1)을 이용한다.

$$VC(x,y) = \cos\theta = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2}\sqrt{\sum_{i=1}^{m} y_i^2}} \qquad \text{(식 1)}$$

정규화된 유클리드 거리는 벡터를 이루고 있는 값들의 분산을 이용하여 (식 2)처럼 구한다.

$$ED(x,y)=\sqrt{\sum_{i=1}^{m}\frac{(x_i-y_i)^2}{\sigma_{i,g}^2}} \quad \text{(식 2)}$$

(식 2)에서 g 는 특정 그룹을 나타낸다. 벡터간의 방향성과 거리를 같이 고려한 새로운 벡터간의 거리 척도 방법은 벡터 상관법과 유클리드 거리에 가중치를 적용하는데 두 벡터 V_i 와 V_t 사이의 새로운 벡터 거리 척도 방법은 (식 3)과 같이 나타낼 수 있다.

$$\begin{aligned} Dist(V_i,V_t) &= GVC(V_i,V_t)+GED(V_i,V_t) \\ GVC(V_i,V_t) &= (1-W_{I,v})\times(1-VC(V_i,V_t)) \\ GED(V_i,V_t) &= W_{I,E}\times ED(V_i,V_t) \end{aligned} \quad \text{(식 3)}$$

(식 3)에서 $i \in I$ 이고, I 는 그룹의 인덱스이다. 또한, $W_{I,v}=VC(G_I,V_t)$ 이며 $W_{I,E}=ED(G_I,V_t)$ 이다. G_I 는 그룹 I 의 특징 벡터를 의미하며, V_i 는 룩업 테이블의 그룹 I 에 속해 있는 i 번째 특징 벡터, V_t 는 추론하고자 하는 시청자의 특징 벡터를 의미한다.

제 1 단계 분류기는 새로운 벡터 거리 척도 방법인 (식 3)을 이용하여 거리 값을 측정한 뒤, 벡터 거리표(VDT)를 만든다. 그림 2 는 제 1 단계 분류기의 측정 방법의 예를 도시하였다.

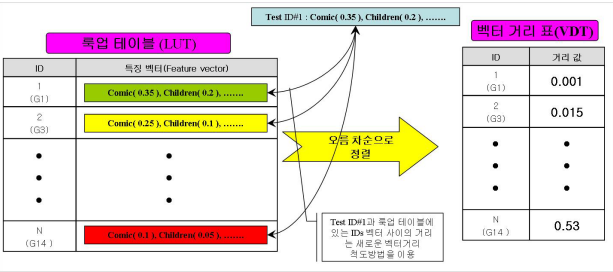

그림 2. 제 1 단계 분류기 측정 방법 예

3.3 제 2 단계 분류기(2nd Stage Classifier)

제 2 단계 분류기는 제 1 단계에서 구한 특징 벡터간의 거리를 이용하여 구한다. 제 2 단계 분류기의 알고리즘은 상위 k 개에 속하는 거리 값을 이용한 가중 거리 k-NN(Weighted-Distance k-Nearest Neighbor) 방법을 이용한다[4]. 가중 거리 k-NN 은 그림 2 의 벡터 거리 표(VDT)내의 상위 k 개의 개수만 고려한다. 그림 2 의 특정 그룹 I 의 가중 거리 k-NN 은 (식 4)와 같다.

$$WDK(I)=\frac{\sum_{i\in I}1/VDT(i)}{\sum_{I=1}^{N}\sum_{j=1}^{k}1/VDT(j,G_I)} \quad \text{(식 4)}$$

(식 4)에서 $i \in I$, I 는 그룹의 인덱스, k 는 k-NN 의 k 값을 의미한다. 또한, $VDT(i)$ 는 k-NN 의 k 개에 속하고 그룹 I 에 속하는 벡터 거리 값을 뜻한다. N 은 그룹의 총 개수를 의미하며, $VDT(j,G_I)$ 는 k-NN 의 k 개에 속하는 G_I 그룹들의 벡터 거리 값이다. (식 4)를 이용하면, VDT 내의 상위 k 개에 속하는 그룹들의 가중 거리 k-NN 표(WDKT)를 만들 수 있다. 그림 3 은 제 2 단계 분류기인 가중 거리 k-NN 을 구하는 방법의 예를 도시하였다.

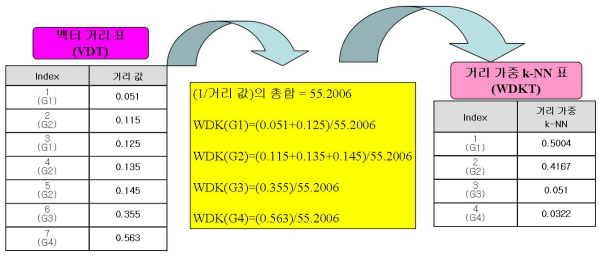

그림 3. 제 2 단계 분류기 측정 방법 예

3.4 제 3 단계 분류기(3rd Stage Classifier)

사용자의 프로파일을 추론하기 위하여 제 1 단계와 제 2 단계를 거치면 요일 별 가중 거리 k-NN 표(WDKT)를 가지게 되었다. 제 3 단계 분류기는 1,2 단계를 거친 요일 별 WDKT 의 최대 값을 이용하여 다수결 원칙 표(MRT)를 작성한다. 제 3 단계 분류기는 다수결 원칙(Majority Rule)과 요일 별 WDKT 내의 최대 값을 이용한 정규화된 다수결 원칙(Normalized Majority Rule)을 이용하여 최종 추론 결과를 낸다. (식 5)는 정규화된 다수결 원칙의 식을 보여주고 있다.

$$NMR(I)=\frac{\max\{WDKT(d)\,|\,d\in D\}}{\sum_{d=1}^{D}\max\{WDKT(d)\,|\,d\in D\}} \quad \text{(식 5)}$$

(식 5)에서 I 는 그룹의 인덱스, D 는 월요일부터 금요일까지를 의미하며, $WDKT(d)$ 는 특정 요일의 WDKT 의 값을 의미한다.

제 3 단계 분류기에서 MRT 와 (식 5)를 이용하면 시청자의 프로파일을 추론할 수 있고, 그림 4 는 이에 대한 예를 보이고 있다.

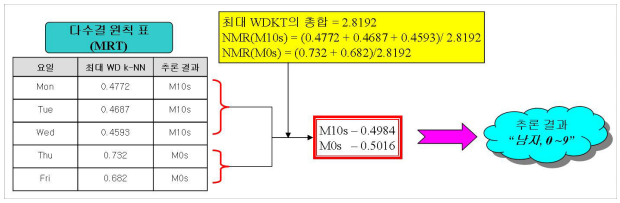

그림 4. 제 3 단계 분류기의 적용 방법 예

IV. 실험 결과 및 표적 광고 시스템 구현

4.1 실험 결과

시청자의 성별 및 연령대를 추론하기 위한 MSC 의 성능을 측정하기 위하여 TV 시청 데이터를 학습 데이터(Training Data)와 테스트 데이터(Test Data) 그룹을 8 개 쌍으로 나누어서 실험을 하였다. 표 4 의 정확도는 8 개 쌍의 정확도를 평균한 결과이다.

성별 및 연령대	정확도(%)		
	VC	ED	MSC
0 ~ 9, M	76.69	71.96	88.21
0 ~ 9, F	67.14	67.86	89.28
10 ~ 19, M	66.89	71.28	89.29
10 ~ 19, F	67.76	68.75	65.79
20 ~ 29, M	60.72	62.95	86.50
20 ~ 29, F	68.18	73.58	78.49
30 ~ 39, M	63.69	72.42	78.80
30 ~ 39, F	63.15	66.88	76.17
40 ~ 49, M	59.82	64.51	86.59
40 ~ 49, F	69.71	64.83	72.25
50 ~ 59, M	54.86	64.58	82.86
50 ~ 59, F	60.86	60.20	67.91
60 ~ M	65.76	67.94	89.77
60 ~ F	56.90	50.86	86.61
계	**64.54**	**66.81**	**80.17**

표 4. 다단계 분류기(MSC)의 실험 결과

표 4 의 실험 결과에서 VC 는 벡터 상관법, ED 는 유클리드 거리, MSC 는 다단계 분류기를 뜻하고, M 은 남자, F 는 여자를 뜻한다. 본 논문에서 제안한 알고리즘이 기존의 방법보다 30% 정확도의 향상을 보이고 있다. 표 4 에서 10 대 여자와 50 대 여자에서 MSC 의 결과가 다른 성별 및 연령대의 결과보다 정확도가 낮은데 이는 특징 벡터가 다른 연령대와 유사한 프로파일을 가지고 있기 때문에 알고리즘의 성능이 비교적 낮은 편이다.

4.2 표적 광고 시스템 구현 결과

본 논문에서 제안된 추론 알고리즘을 이용하여 표적 광고 프로토타입 시스템 구현 결과를 제시한다. 표적 광고 프로토타입 시스템을 위한 광고 콘텐츠는 NGTV 에서 무료로 제공되는 콘텐츠를 이용하였고, 광고 콘텐츠들은 성별 및 연령대에 따라 임의로 할당하였다. 그림 5 의 프로토타입 시스템은 시청자가 성별과 연령대를 주어지지 않았을 때, 본 논문에서 제안하는 MSC 알고리즘을 이용하여 서버에서 광고 콘텐츠를 전송한 결과를 보여주고 있다.

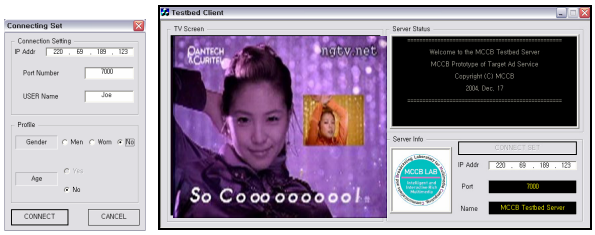

그림 5. 표적 광고 프로토타입 시스템 구현 결과

V. 결론

본 논문은 디지털 방송의 다채널 시청환경에서 시청자의 TV 시청 데이터를 이용하여 표적 광고 서비스에 관한 연구로서, 시청자의 TV 시청 데이터를 바탕으로 특징 벡터를 추출하여 다단계 분류기(Multistage Classifier)를 이용하여 시청자의 성별 및 연령대에 대한 프로파일을 기반으로 한 표적광고 프로토타입 시스템을 제안하였다. 제안된 표적 광고 시스템을 통하여 시청자들은 자신의 성별과 연령대에 맞는 광고를 볼 수 있으며, 콘텐츠 공급자는 광고로 인한 수익 증대를 꾀할 수 있고, 광고주는 회사가 표적으로 삼고 있는 고객들에게 광고를 제공함으로써 보다 효과적인 광고 효과를 볼 수 있다.

참고문헌

[1] Dimitrios Katsaros and Yannis Manolopoulos, “Broadcast program generation for webcasting”, Source Data & Knowledge Engineering, Vol 49 , Issue 1, pp. 1~21, April 2004

[2] Theodore Bozios, George Lekakos, Victoria Skoularidou and Kostas Chorianopoulos, "Advanced Techniques for Personalised Advertising in a Digital TV Environment: The iMEDIA System.", Proceedings of the E-business and E-work Conference, 17-19 October, Venice, Italy, pp. 1025-1031, 2001.

[3] C. Shahabi, A. Faisal, F.B. Kashani and J. Faruque, " INSITE: A Tool for interpreting Users", Proceeding of Interaction with a Web Space. Volume00, pp: 635-638.

[4] Wei Yuan, Juan Liu, and Huai-Bei Zhou, ”An Improved KNN Method and Its Application To Tumor Diagnosis”, Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, August 2004.