지능형 개인화 EPG를 위한 프로그램 정보 장르 분류

송진석
고려대학교 컴퓨터정보통신대학원 미디어공학과
e-mail:luckyfrog@korea.ac.kr

Classification of Program Information Genre for Intelligent Personalized EPG

Jinseok Song
Dept. of Media Engineering.
Graduate School of Computer and Information Technology,
Korea University.

요 약

국내에서 디지털 방송 상용화에 성공하고 전송 모델 또한 다양화됨에 따라 사용자는 다양한 형식으로 다수의 방송 프로그램을 접할 수 있게 되었다. 이에 대한 효율적인 프로그램 관리를 위한 EPG(Electronic Program Guide) 서비스가 현재 제공되거나 개발 중이다. 지능형 개인화 EPG는 디지털 방송 스트림이 수신되는 환경에서 사용자와 방송 수신기의 지능적인 매개체로서 운영되며 본 연구는 기존 프로그램 정보에 대한 장르를 학습하고 새로운 프로그램 정보가 입력될 경우 올바르게 장르를 분류할 수 있도록 기계학습 기법이 사용되었다.

1. 서론

처음으로 DVB(Digital Video Broadcasting)-MHP(Multimedia Home Platform) 표준[1]의 디지털 위성 방송 상용화 서비스가 성공하고 현재 채널 180개의 채널과 시간당 상영되고 있는 방송 프로그램은 80개, 하루에 평균 1400개의 프로그램이 방송되고 있다. 또한 방송 수신기가 디지털방송 위성 STB, 케이블 STB, IPTV 수신기, DMB (Digital Multimedia Broadcasting) 수신기 등 다양화 되고 지상파 방송, 케이블 방송, VOD 서비스, Flash, UCC 등 최근 서비스 형태도 다양화 되고 있다.

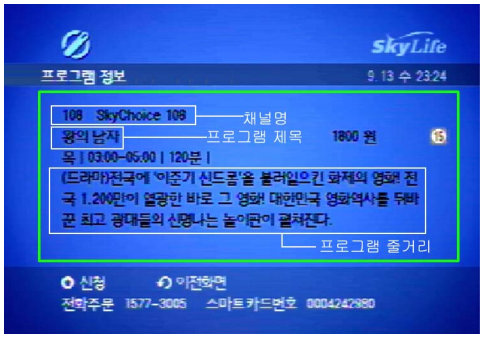
이러한 다수 방송프로그램의 효율적인 관리를 위한 EPG의 필요성이 인식되고 있으며 현재는 개인에게 능동적으로 방송프로그램을 제안해 줄 수 있는 '지능형 개인화 EPG 서비스'에 대한 많은 연구 진행되고 있다.[3]

기존에 제안된 지능형 TV 프로그램 가이드의 경우 장르에 대한 가중치가 가장 높으며 사용자가 원하는 프로그램을 장르를 기반으로 다양하게 제공하기 위해서는 장르에 대한 정확한 구분이 필요하다. 또한 장르가 구분되지 않은 프로그램 정보가 입력될 경우 자동으로 장르 분류를 할 수 있어야 하며 이를 위해 기계학습 기법이 사용되었다.[6]

본 논문의 구성은 다음과 같다. 2장 EPG의 활용에서는 방송 프로그램 정보를 어떻게 추출하고 장르 분류를 위해 이용할 것인지 제시하였고, 3장 프로그램 정보 장르 분류에서는 전체적인 시스템 흐름을 설명하였다. 4장 Input Data 생성 과정과 5장 기계학습 알고리즘 적용에서는 장르별 색인 과정과 학습 과정 추후 장르가 분류되지 않은 새로운 데이터에 대해 분류가 가능하도록 기계학습에 대한 설명을 하였다. 6장에서는 제안한 모델의 실험 결과 도출 후 7장에서 논문의 결론을 제시하였다.

2. EPG (Electronic Program Guide)의 활용

현재 디지털방송에서는 방송신호와 방송 가이드 정보를 다중으로 보내고 있고 시청자는 별도의 수신기를 가지고 가이드 정보를 수신하도록 되어 있다.



(그림 1) EPG 프로그램 정보

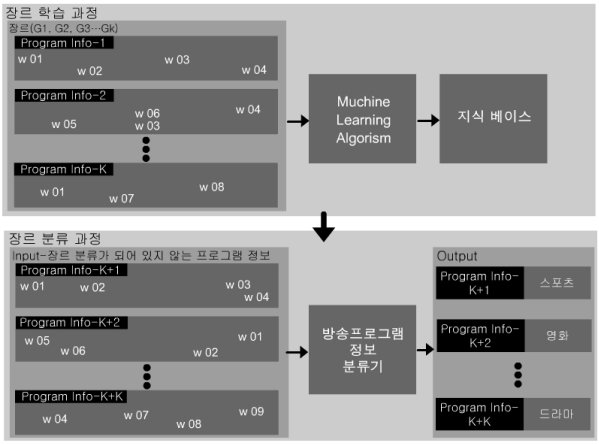
방송 가이드 정보를 SI(Service Information)이라 하며[2] 이는 EPG라는 Interface를 통해 시청자가 원할 때 TV로 시청할 수 있다. (그림1)은 실제 한국 디지털위성방송에서 서비스하고 있는 EPG 프로그램 정보가 출력되는 화면이며 장르 분류를 위해 <표1>과 같은 정보를 추출하여 학습데이터로 활용하였다.

3. 프로그램 정보 장르 분류

본 연구에서 세분화된 장르 구분을 위해 '유네스코 분류법'이라는 프로그램 분류 방법 중 장르의 속성 값을 사용하였다.[7] 총 11개의 장르를 분류하여 사용하고 있는데 이중 사용자가 가장 많은 정보를 보유하고 있는 5개의 장르를 분류에 사용하였다.

채널명	예) 201번 KBS, 202번 MBC, 203번 SBS, 301번 OCN, 642번 MBC GAME
프로그램 제목	예) 왕의 남자, 슈퍼맨, 강호동의 천생연분, K-1 월드 그랑프리
프로그램 줄거리	출연진 : 이형민,이경희 15년간 한결같이 사랑하는 여자 은환(공효진)의 아름답고 따뜻한 사랑 이야기가 줄거리다.

<표 1> Sample Data에 추출되는 정보의 예



(그림 2) 장르 분류를 위한 전체 시스템 흐름도

<표1>은 장르 분류를 위해 디지털 방송에서 서비스되고 있는 예제이며 SI정보 중 자연어만을 추출하여 기계학습을 위한 Sample Data로 활용이 되었다.

(그림2)는 프로그램 정보 분류를 위한 전체 흐름도이며 크게 장르 학습 과정과 장르 분류 과정으로 나뉜다. Program Info-1, Program Info-2, …,Program Info-K까지의 Sample Data는 학습을 위해 이미 장르 분류가 되어있는 데이터들이고 Program Info-K+1, Program Info-K+2, …,Program Info-K+K는 장르 분류가 되어 있지 않은 데이터들이다. 이러한 기계학습 방법은 알려진 예에 대한 학습을 통해 새로운 결과를 예측하려는 Supervised Learning 시스템에 기초하며 다음 장은 기계학습 알고리즘에 입력될 Input Data를 생성하는 과정에 대한 설명이다.

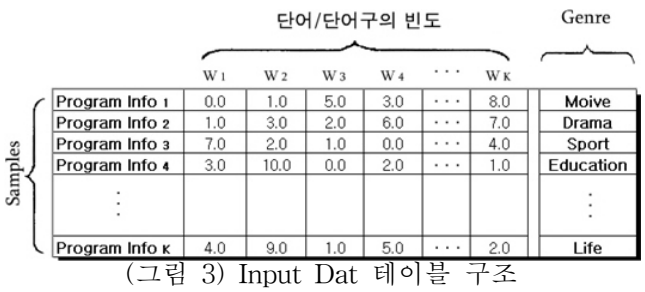
4. Input Data 생성 과정

4.1 장르별 색인

장르별 특징을 추출하는 과정으로 장르의 속성을 표현하는 단어(구)를 뽑아 해당 장르의 내용을 대표하는 과정이다.

순서는 데이터에 있는 채널명, 프로그램 제목, 프로그램 줄거리를 품사 태깅 작업이 있은 후 단어(구)들을 추출해 낸다. 이후 단어의 빈도(term frequency)가 높은 단어(구)들을 추출하여 빈도별 내림차순으로 정렬하였다. 이중 타 장르에 비해 해당 장르에서 빈도가 높은 용어들을 추출하여 장르를 대표하는 색인어로 사용하며 이후 벡터화 작업을 진행하게 된다.

4.2 Input Data 테이블 생성



(그림 3) Input Dat 테이블 구조

(그림3)에서 Program Info-1, Program Info-2, …Program Info-K는 <표1>에서 예로 제시된 각각의 프로그램 제목들과 매치된다. W 01, W 02, W 03, … W k는 장르별 색인 리스트에 포함되어 있는

단어(구)의 출현 빈도가 벡터 값으로 입력되도록 되어 있으며 Genre에 입력된 Movie, Drama 등은 학습을 위해 이미 분류된 해당 장르의 값이다.

5. 기계학습 알고리즘 적용

본 연구에서는 방송 프로그램 정보에서 추출되고 구성된 하나의 Input Data를 정확률 높고 사용자 환경에 적합한 학습 알고리즘을 구분해 내기 위해 베이지안(Bayesian)과 신경망(Neural Network), 결정트리(Decision Tree) 학습 알고리즘으로 비교 실험을 하였다.[8][9]

베이지안 학습 알고리즘은 확률 이론을 기계학습에 적용한 기법으로 구현이 용이하다는 장점으로 인해 기존 문서분류시스템에 많이 사용되어 왔다. 베이지안은 많은 양의 데이터를 입력했을 때 정확률이 다른 학습기법 보다 더 높다고 알려져 있으며 기존 문서분류에 대한 비교실험에 많이 사용되어 왔다.

신경망 학습 알고리즘은 베이지안 학습 알고리즘과 함께 기계학습 기법에 가장 많이 쓰이고 있으며 Input Data에 Noise가 있어도 비교적 견고하게 동작하는 편으로 알려져 있다. 네트워크를 통해 feeding된 에러를 backward 시키는 방법인 역전파 알고리즘(Back-Propagation Algorithm)을 사용하여 학습하였으며 Layer 의 개수 전체 3개, Hidden Layer 1개로 구성 하였다. 각 Layer의 Unit 개수는 Input Layer는 100개, Hidden Layer는 53개, Output Layer 5개로서 이는 5개 장르 (영화, 드라마, 스포츠, 교육, 생활정보)를 예측하여 Output Data로 출력하게 된다.

결정트리 기법은 수집된 sample의 데이터를 분석하여 장르 분류를 위한 규칙을 트리 모양으로 구성해 낸다. 이 기법은 분류의 근거를 확인할 수 있기 때문에 변수간 관계와 영향을 이해하기가 쉽다. 생성되는 의사결정트리의 depth는 50으로 제한을 두었고 복수 의사 결정트리 기법을 사용하였다.

6. 실험 및 평가

현재 위성방송에서 전송되고 있는 180개 채널의 프로그램 정보 중 4주 분량의 데이터를 추출하였고 이중 가장 많은 데이터를 포함하고 있는 5개의 장르(영화, 드라마, 스포츠, 교육, 생활정보)에 해당하는 정보들을 수집하였다.

수집된 Sample 개수는 총928개이며 각각 장르별로 수집된 Sample의 개수는 영화 212개, 드라마 206개, 스포츠 170개, 교육 170개, 생활정보 170개이다.

6.1 실험 기법

타 실험에서 보면 기본적으로 학습데이터는 70%, 테스트데이터는 30%를 사용하고 있지만 본 실험에 수집된 데이터는 상대적으로 적기 때문에 5x2 Cross Validation 테스트 기법을 사용하였다.

$$
\begin{aligned}
T_1 &= X_1^{(1)} \quad V_1 = X_1^{(2)} \\
T_2 &= X_1^{(2)} \quad V_2 = X_1^{(1)} \\
T_3 &= X_2^{(1)} \quad V_3 = X_2^{(2)} \\
T_4 &= X_2^{(2)} \quad V_4 = X_2^{(1)} \\
&\vdots \\
T_9 &= X_5^{(1)} \quad V_9 = X_5^{(2)} \\
T_{10} &= X_5^{(2)} \quad V_{10} = X_5^{(1)}
\end{aligned}
$$

(수식 1) 5x2 Cross Validation 기법 적용 방법[8]

(수식1) 중 $X_i^{(j)}$ 는 fold를 i로 표현하고 fold i의 절반은 j로 표현한다. 즉$X_1^{(1)}$ 와 $X_1^{(2)}$ 에 random하게 혼합된 928개의 데이터가 절반씩 나눈다. 5개의 fold 안에는 각각 다른 순서를 가진 928개의 데이터들이 포함되어 있으며 총 4640개의 데이터가 Validation Data로 사용되었다. Validation Data의 경우는 첫 번째 fold를 포함하는 T_1과 T_2 중 T_1의 sample들은 V_2의 sample로 사용하고 T_2의 sample들은 V_1의 sample로 교차되어 사용된다.

6.2 성능 평가

	전체	1회	2회	3회	4회	5회
영화 1	530	83.02%	73.58%	75.47%	71.70%	66.04%
영화 2	530	68.87%	68.87%	80.19%	82.08%	68.87%
드라마 1	515	58.25%	71.84%	81.55%	75.73%	78.64%
드라마 2	515	52.43%	67.96%	71.84%	75.73%	63.11%
스포츠 1	425	85.88%	88.24%	87.06%	81.18%	83.53%
스포츠 2	425	84.71%	82.35%	81.18%	82.35%	92.94%
교육 1	425	76.47%	91.76%	85.88%	84.71%	80.00%
교육 2	425	88.24%	95.29%	75.29%	84.71%	76.47%
생활 1	425	76.47%	85.88%	88.24%	89.41%	77.65%
생활 2	425	78.82%	70.59%	85.88%	83.53%	82.35%
정확률	4640	75.32%	79.64%	81.26%	81.11%	76.96%

<표 2> 베이지안 학습을 통한 장르 분류 결과

	전체	1회	2회	3회	4회	5회
영화 1	530	92.45%	90.57%	88.68%	80.19%	82.08%
영화 2	530	91.51%	80.19%	93.40%	88.68%	81.13%
드라마 1	515	76.70%	72.82%	72.82%	84.47%	80.58%
드라마 2	515	88.35%	78.64%	78.64%	82.52%	79.61%
스포츠 1	425	87.06%	68.24%	67.06%	95.29%	89.41%
스포츠 2	425	97.65%	92.94%	82.35%	58.82%	91.76%
교육 1	425	74.12%	97.65%	97.65%	76.47%	92.94%
교육 2	425	67.06%	100%	82.35%	84.71%	90.59%
생활 1	425	90.59%	85.88%	85.88%	89.41%	82.35%
생활 2	425	83.53%	81.18%	85.88%	84.71%	91.76%
정확률	4640	84.90%	84.81%	83.47%	82.53%	86.22%

<표 3> 신경망 학습을 통한 장르 분류 결과

	전체	1회	2회	3회	4회	5회
영화 1	530	70.75%	79.25%	81.13%	75.47%	83.02%
영화 2	530	86.79%	75.47%	81.13%	78.30%	72.64%
드라마 1	515	76.70%	60.19%	75.73%	73.79%	69.90%
드라마 2	515	59.22%	70.87%	67.96%	70.87%	76.70%
스포츠 1	425	91.76%	91.76%	91.76%	98.82%	98.82%
스포츠 2	425	100%	100%	100%	92.94%	92.94%
교육 1	425	98.82%	100%	96.47%	84.71%	92.94%
교육 2	425	78.82%	100%	90.59%	95.29%	96.47%
생활 1	425	92.94%	87.06%	96.47%	96.47%	90.59%
생활 2	425	94.12%	76.47%	88.24%	90.59%	91.76%
정확률	4640	84.99%	84.11%	86.95%	85.73%	86.58%

<표 4> 결정트리 학습을 통한 장르 분류 결과

위 표는 베이지안, 신경망, 결정트리 학습을 통한 장르 분류 실험 결과이다. 두 번째 열의 숫자는 분류 실험을 위한 test data 개수를 나타내고 세 번째 열부터는 횟수마다 해당 학습 알고리즘을 통해 분류된 test data의 정확률을 나타낸다. 정확률에 대한 계산은 아래 (수식2)와 같이 계산되었다.

$$\text{장르분류의 정확률}(\%)=\frac{\text{정확하게 장르분류된 프로그램 개수}}{\text{입력된 전체 프로그램 개수}}\ X100(\%)$$

(수식 2) 장르 분류의 정확률 추출을 위한 식

본 연구에서는 패턴 정확률만을 가지고 성능을 측정하였으며 패턴 정확률은 적용된 장르 분류 시스템이 어느 정도 정확하게 장르를 분류했는지를 나타낸다.

베이지안	신경망	결정트리
78.86%	84.39%	85.67%

<표 4> 기계학습 알고리즘 비교 실험 결과

7. 결론

본 연구에서는 프로그램 정보를 통해 추출된 동일한 sample data를 각각의 학습 알고리즘에 적용하여 비교실험 하였으며 실험 결과를 보면 베이지안 알고리즘이 가장 낮은 성능을 보여주었고 신경망 알고리즘 보다는 결정트리 알고리즘이 약간은 높은 성능을 보여주었다.

본 연구에서 사용된 실험기법은 5x2 Cross - Validation 기법을 사용하여 Training과 Validation data가 반반으로 나누었기 때문에 타 실험기법에 비해 상대적으로 장르 분류에 대한 정확률이 떨어지며 7:3의 비율로 실험할 경우 정확률은 더 높아질 것으로 예상한다.

앞으로 이러한 기법을 통해 장르의 속성 분류와 내용의 속성 분류를 정합하여 사용한다면 사용자가 사용하는 방송 프로그램 정보의 장르와 세부장르에 대해 향상된 정보를 제공할 수 있을 것이며 이는 사용자가 시청할 방송 프로그램 예측의 정확률을 높이는데 중요한 데이터로 활용될 것이다.

참고문헌

[1] Piesing J. “The DVB multimedia home platform (MHP) and related specifications”, Vol. 94, No. 1, pp. 237-247, 2006.

[2] “Specification for Service Information (SI) in DVB systems”, A038, Rev. 1, 2000.

[3] 류지웅, 김문철, 남제호, 강경옥, 김진웅, “사용자 선호도 기반 지능형 프로그램 가이드”, 한국방송공학회 제7권, 제2호, pp. 153-167, 2002.

[4] Joachims, T., "Text categorization with support vector machines: learning with many relevant feature", In Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137-142, 1998.

[5] L, Ardissono. “User Modeling and Recommendation Techniques for personalized Electronic Program Guides, In Personalized Digital Television, Human-Computer Interaction Series”, Vol. 6, 2004.

[6] 김상범, “범주간의 상호관계를 고려한 자동 문서 범주화의 개선”, 고려대학교 컴퓨터학과 전산학석사 학위논문, 1999.

[7] 이재훈, 정문렬, “속성 값들의 연관관계를 이용한 EPG User Interface의 설계”, 한국방송공학회 학술대회지, pp. 87-90, 2003.

[8] Ethem Alpaydin, Introduction to Machine Learning, MIT Press, 2004.

[9] T. Mitchel, Machine Learning, McGraw, 1997.