2007년도 대한전기학회 하계학술대회 논문집 2007. 7. 18 - 20

시간 확장형 베이시안 네트워크를 이용한 TV 시청자 채널 추천 방법

김지나* 임태범** 윤경로*
건국대학교 컴퓨터공학과*, 전자부품연구원**

TV Channel Recommendation Method Using Temporally Extended Bayesian Network

Jina Kim*, Tae-Beom Lim**, Kyoungro Yoon*
Department of Computer Science and Engineering Konkuk University*, Korea Eletronics Technology Institute**

Abstract - 최근 디지털 TV방송서비스의 보급으로 채널의 수와 그에 따른 프로그램의 수가 많아짐에 따라 시청자는 모든 프로그램의 정보를 미리 알고 있는 것이 힘들어 졌다. 모든 채널과 방송 프로그램을 탐색하고 자신의 취향에 맞는 프로그램을 찾아보기 어려워진 문제를 해결하고자 영화, 상품 등의 분야에 국한되었던 추천연구 분야도 TV채널 까지 확대할 필요가 있다. 본 논문에서는 사용자의 TV시청 기록을 분석하여 사용자 프로파일 테이블을 구성하고, 베이시안 네트워크와 시계열 분석 이론을 접목하여 추천엔진을 구현하는 TV채널 추천 엔진을 제안한다.

1. 서 론

최근 케이블 TV와 지상파, 위성 디지털 방송서비스의 보급으로 지상파 TV방송 서비스 환경보다 늘어난 채널과 프로그램의 다양화로 시청자들을 만족시키고 있다. 하지만 너무 많은 채널의 수로 인해 오히려 사용자가 원하는 프로그램을 제때 찾아 시청하기 힘들어지는 현상도 생겨나고 있다. 과거에는 영화, 상품, 음악 등을 사용자에게 추천하는 '추천방법' 연구가 TV채널 추천 영역으로 확대될 필요가 생기는 것이다. 추천이란 다른 의미로 사용자가 소비할 아이템을 예측 하는 것과 같은 의미로, 본 논문에서는 '베이시안 네트워크'를 기반으로 예측을 한다. 베이시안 네트워크란 사전에 일어난 일을 토대로 사후의 확률을 추론하는 방법으로, 사용자가 시청했던 기록을 분석하여 다음에 무엇을 시청할 지를 예측하는 방법이다. 그러나 TV시청은 최근의 시청 경험이 과거의 시청 경험보다 사용자의 채널 선택에 더 큰 영향을 미치는 경향이 있으므로, 시간 축에 따라 데이터의 비중을 달리 할 필요가 있어, 시간 축에 따라 데이터를 각기 다른 의미로 분석하는 '시계열 분석 이론'을 함께 사용하였다. 본 논문에서는 베이시안 이론과 시계열을 접목한 시간 확장형 베이시안 네트워크를 제안하고 이를 기반으로 방송프로그램을 추천하여 본다.

2. 본 론

2.1 TV프로그램의 모델링

사용자의 TV 시청기록을 기반으로 사용자 프로파일을 구성하기 위하여, 사용자의 프로그램 시청기록을 모델링하여야 한다. 이때, 하나의 TV 프로그램은 여러 개의 factor로 나뉘어져 모델링된다. 각각의 factor는 시청자가 프로그램을 선택하는 기준이 되는 구체적인 항목들로 다음과 같이 방송사, 장르, 세부장르, 출연진으로 나뉜다.

-채널: 프로그램을 방영하는 방송사이다. 만약 같은 시간대에 방송 3사에서 모두 뉴스를 방영하고 있다고 하자. 뉴스에서 보도하는 내용들이 대부분 비슷하다고 한다면 사용자에 따라서 특정 방송사의 견해를 더 공감할 수 있고, 그러한 경우 그 채널의 뉴스를 시청할 확률이 높아진다. 즉 모든 환경이 같다고 할 때 채널 (또는 방송사)이 시청 기준이 될 수 있으므로 방송사가 factor가 되게 된다.

-장르: 방송되는 프로그램의 분야를 말한다. 크게는 뉴스, 드라마, 영화, 쇼/오락, 시사다큐. 교양 등으로 나뉜다. 사용자가 드라마를 보는 것을 좋아한다면 추천 대상에서도 드라마가 상위에 기록될 것이다.

-세부장르: 장르가 크게 드라마라 할지라도 멜로드라마, 홈드라마, 사극드라마, 의학드라마 등의 여러 주제로 드라마를 나눌 수 있다. TV프로그램의 주제가 다양화됨에 따라 더 많은 종류의 세부장르가 생기게 된다. 세부장르는 장르라는 상위노드에 속하는 하위노드로 볼 수 있지만, 사용자가 스포츠중계를 즐겨 본다면 스포츠를 세부장르로 하는 드라마를 볼 확률도 높아진다는 예에서 볼 수 있듯이 다른 장르라 할지라도 같은 세부장르를 가질 수 있다. 즉 장르에 국한되지 않는 고유의 factor로 세부장르를 나눌 수 있다.

-출연진: 프로그램의 선택에서 또한 중요하게 여겨지는 항목이 출연진이다. 자신이 좋아하는 출연진이 출연하는 방송이라면 장르나 세부장르 등의 사전 정보 없이도 프로그램을 선택할 확률이 높은 중요한 선택의 기준이 된다.

본 논문에서는 사용자 프로파일을 작성하기 위해 시청기록 테이블을 작성하고 이때 세부 항목은 채널, 장르, 세부장르, 출연진1, 출연진2의 5개 항목으로 한다. 출연진을 두 개로 나누는 이유는 어떤 방송은 출연진이 거의 나오지 않으며 또한 채널 선택에 큰 영향을 미치지 않지만, 드라마나 쇼 프로그램 같은 경우는 많은 출연자가 등장하기 때문에 하나의 출연인물 만으로 모두 반영하기 힘들기 때문이다. 이런 이유로 본 논문의 실험에서는 대표하는 출연진 두 명을 출연진1과 출연진2로 구성한다.

2.2 베이시안 알고리즘과 시계열의 적용

시청 기록의 각각의 항목들(방송사, 장르 등 시청 기록에서 뽑아낸 factor)을 가지고 추천결과를 찾아내는 과정은 <그림1>의 식1로 나타낼 수 있다. F는 추천할 항목(Factor)으로 만약 시청할 프로그램의 방송사를 계산한다면, F는 방송사가 되고 f_j는 방송사의 종류인 MBC, KBS1등이 된다. 그리고 a_1, a_2, ...,a_n은 입력 값으로 들어갈 시청기록 데이터의 모든 세부 항목들(방송사에서 출연진2까지)이 된다. 또한 각각의 a_i가 서로 독립적이라는 조건 하에 P는 식 2로 정의할 수 있고 식1과 식2를 조합하면 식3과 같은 최종 식을 얻을 수 있다.

$$Rec(F) = \arg max F_{j \in F} = P(f_j) P(a_1, a_2, ..., a_n | f_j) \quad \text{-식1}$$

$$P(a_1, ..., a_n | f_j) = \prod_i P(a_i | f_j) \quad \text{-식2}$$

$$F = \arg max_{f_{f \in F}} = P(f_j) \prod_i P(a_i | f_j) \quad \text{-식3}$$

<그림 1> 베이시안 확률 추론식

식 3을 계산하기 위해서 해당 항목이 추천될 확률인 $P(a_i|f_j)$의 값을 알아야 하는데 이는 $\frac{n_c}{n}$정도로 생각할 수 있다. 여기서 n은 recommend테이블에 추천할 항목이 있는 f_i의 개수이고, n_c는 f_i를 만족하는 a_i 즉, 전체 데이터 셋 중에서 특정 조건을 만족하는(추천을 위해 사용자의 기록과 추천 대상이 서로 일치하는) 셋의 개수가 된다. 하지만 $\frac{n_c}{n}$는 데이터의 양이 많을수록 정확도가 커지는데 실험에서는 데이터 셋의 개수를 충분할 만큼 많은 수로 실험하기 힘들고, 데이터의 개수가 작은 만큼 수치의 정확도가 떨어지게 된다. 이 문제를 해결하기 위해 Machine Learning 이론 중에서 M-Estimation을 적용하여 $P(a_i|f_j) = \frac{n_c + mp}{n+m}$라는 식으로 바꾸어 사용한다. 여기에서 p는 데이터 셋의 환경에 따라 바꾸어 적용하는 일종의 '추정치' 로 정확한 수치는 입력하는 데이터의 수에 따라 다른 수치가 된다.

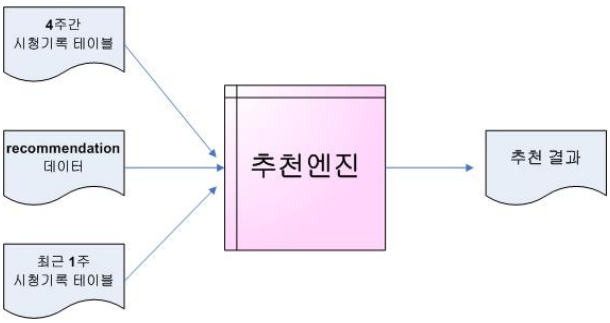
방송을 시청할 때 같은 시간 같은 방송을 반복적으로 시청하는 경우도 있지만 보통의 환경에서 시청할 방송을 선택할 경우 사용자는 자신이 최근에 시청한 방송에 대한 기억이 선택에 큰 영향을 미치게 된다. 오늘 시청할 방송을 선택한다면 같은 시간대의 시청 기억이 1개월 전 시청 기억보다는 1주일 전 시청 기억이 더 큰 영향을 미치고 1주일 전 시청 기억 보다는 바로 어제 시청한 기억이 오늘 시청할 방송 선택에 더 큰 영향을 미친다는 것이다. 이 내용을 베이시안에 적용한다면 시계열로 베이시안을 확장하는 것이 된다. 시청한 기록을 시간에 따라 각각 다른 정도로 달리 적용하는 가중치를 반영하는 것이다. 이것을 식으로 나타내면

$P(a_i|f_j)=\frac{(\alpha\times n_c)+mp}{n+m}$이 된다. 이 식에서 α는 가중치로, 테이블에서 추천 조건을 만족하는 프로그램을 찾아 빈도수를 시간 축에 따라 다르게 가중치를 적용하여 더한다는 의미이다. 즉, 조건을 만족하는 데이터 셋을 찾으면 빈도를 높이되, 가장 최근 1주일간은 높은 가중치를 적용하여 더하고 추천 날짜에서 멀어질수록 작은 가중치를 적용하여 더하는 것이다.

2.3 시청기록 데이터의 가공

실험을 위한 데이터는 닐슨에서 제공하는 시청률 자료 중 2007년 1월 1일부터 2007년 2월 28일 까지 2개월간의 약 8800개의 데이터를 사용하였다. 시청 기록은 당일 2:00부터 25:59 까지 시청 기록을 1분 단위씩 1440개의 방송사 앞 글자 이니셜로 되어 있다. 먼저 시청 기록 중에서 같은 특정 사용자의 두 달간의 기록을 추출해 내고, 그 기록들 중 실험하고자 하는 시간대의 기록만을 따로 추출하는 과정을 거친다. 만약 9:00~9:59 사이의 시청 기록이라면 1분단위로 60개의 기록이 저장될 것이다. 이 기록을 10분단위로 묶어서 기록한다. 즉 한 시간 동안 시청한 6개의 프로그램을 저장하는 것이다.

이렇게 얻은 프로그램 기록은 그 프로그램의 방송사, 장르, 세부장르, 출연진이 무엇인지 기록하는 세부항목 테이블을 작성해야 한다. 세부 항목은 EPG.com에서 제공하는 해당 날자의 채널과 프로그램 정보를 이용하여 찾아내었다. 파일에서 정보를 읽어 들인 다음, 각 프로그램의 데이터에 시청 시간대와 날자 별로 프로그램 세부 항목을 넣은 데이터를 작성한다. 시청 기록 테이블은 일주일간 시청 기록 테이블을 한 셋으로 실험하고자 하는 주간의 데이터를 1주일 단위로 구분해서 저장한다. 예를 들어, 9:10~9:29 사이의 30분 동안 시청한 4주간 기록을 이용한다면 5(세부항목=방송사, 장르, 세부장르, 출연진1, 출연진2)×1(일주일단위)×3(10분단위 3개)×4(4주간)=60 총 420개의 데이터를 이용하는 것이 된다. 이렇게 4주간의 시청 기록을 table로 저장한다. 다음으로 필요한 기록은 recommendation테이블로 1주일간 시청한 기록이 table로 주어지면 그 1주일 바로 다음날은 어떤 방송을 보았느냐가 recommendation테이블이 된다. 즉, 1주일간 시청한 기록이 조건이 되고 그 다음날 시청한 기록이 결과가 되는 것으로 '1주일간 이런 방송을 시청했을 때 사용자의 선호도에 의해 다음날은 이 방송을 본 것이다' 라는 가정을 하는 것이다. 마지막으로 필요한 테이블은 추천 날짜 바로 직전 1주일간 시청 기록이다. 앞의 table 시청 기록에 의해 recommendation이라는 시청 결과가 주어졌으니 최근 1주 시청 기록을 엔진에 넣으면 추천 결과가 나오는 원리이다. 가장 최근 시청 기록을 data라 하고 추천 결과파일을 result라고 하면 'table:recommd=data:result' 라고 표현할 수 있다.



<그림 2> 추천 엔진의 동작 과정

2.4 추천 엔진의 구현

사용자의 시청 기록이 1주부터 6주까지 있다. 각 주의 매일 매일 특정 시간을 단위로(실험에서 10분) 예측 시간을 포함하여 뒤로 4개의 데이터가 있다. 1주차의 데이터를 기반으로 2주의 결과, 2주차의 데이터를 기반으로 3주차의 결과, 3주차의 데이터를 기반으로 4주차의 결과가 나오고, 4주차의 데이터를 기반으로 5주의 결과를 예측했다고 가정한다. 즉, 5주차의 결과를 보고, 이러한 결과를 발생시키는 앞의 조건 들이 어떻게 되는지 1주부터 4주까지의 기록을 분석하는 방식이다. 4주의 데이터를 기반으로 P(방송사|드라마)를 구하기 위하여 1주부터 4주까지의 데이터에서 5주차와 방송사가 모든 요일에 일치하는 시간대 자료를 구한다. (월~일요일까지 7개 방송사의 조합이 동일한 자료) 그 중에서 5주차의 실제 시청한 결과 중에 장르가 '드라마' 인 자료를 찾는다.

예를 들어, 화요일 저녁 9시 3분에 추천을 하기 위하여, 그 전주 화요일~월요일까지 시청기록을 보유하는 9시부터 9시 40분까지의 10분 단위의 4개의 데이터가 있다. 모두 4주의 데이터를 가지고 실험하므로, 전체 16개의 데이터가 있는 것이다. 그중 데이터의 다음 주 화요일 저녁 9시~9시 40분 사이의 10분 간격으로 자른 4개의 데이터 중 장르가 '드라마'인 데이터를 확인한다. (개수가 16개중 n개가 된다.) 선택된 n개 중에서 그 전주 데이터의 방송사 정보가 추천하고자 하는 바로 지난주의 4개의 데이터 중 최소 어느 하나와 일치하는 데이터를 찾는다. (n개 중에서 n_c개가 된다.) 이렇게 매칭 되는 데이터를 $n_c(i)$라고 하고 각 주간마다 가중치$\alpha(i)$를 적용하면 $P(\text{방송사}|\text{드라마})=\frac{\sum_i[\alpha(i)\times n_c(i)]+mp}{n+m}$ 라는 식이 된다. 실제로 계산한 수치가 n은 10, $\sum(\alpha i \times n_c i)$는 5.8 이고 드라마가가 속한 장르의 종류는 6개, 하나의 데이터 셋의 총 개수는 5×7=35였다고 하면 $P(\text{방송사}|\text{드라마})=\frac{5.8+\frac{35}{6}}{10+35}=\frac{689}{2700}=0.25518$ 라는 결과가 나온다. 이렇게 모든 항목에 대하여 계산하여 P(방송사|드라마), P(장르|드라마), P(세부장르|드라마), P(출연진1|드라마), P(출연진2|드라마)를 각각 구하여 모두 곱하면 추천 엔진이 드라마를 추천할 확률인 P(드라마) 가 나온다. 이 수치를 정렬하여 가장 큰 값을 가진 항목을 추천결과로 출력하는 것이다.

	월	화	수	목	금	토	일
방송사	MBC	SBS	SBS	MBC	KBS2	KBS1	MBC
장르	드라마	드라마	드라마	드라마	시사/교양	교양정보	드라마
세부장르	역사/사극	홈드라마	트렌디	트렌디	교양일반	언론	의학
출연진1	송일국	김동완	이서진	세븐	황정민	없음	김명민
출연진2	한혜진	한은정	김정은	허이재	없음	없음	이선균

<표 1> 1주일단위 시청 기록 데이터 한 셋의 예

3. 결 론

추천 성능의 측정은 추천 엔진의 결과와 그 사용자의 실제 시청 기록을 비교하여 측정하였다. 세부 항목 별로 가장 높은 수치의 항목을 조합하여 일치하는 프로그램을 찾아서 추천하는 것인데 모두 일치하는 방송이 없을 때는 2순위 항목으로 조합하였다. 실험 결과 방송사, 장르, 세부장르는 비교적 추천의 성공률이 높았지만 출연진의 경우는 실험 데이터의 패턴에 따라 성공률의 차이가 많음을 알 수 있었다. 이는 한 프로그램에 출연진의 수가 많은 장르와 뉴스나 시사프로그램 등 출연진의 수가 많지 않거나 출연진이 프로그램 선택에 영향을 많이 끼치지 않는 장르와의 차이에서 비롯된 것으로 보인다. 이러한 문제는 향후 연구에서 장르별로 다른 단계의 베이시안 네트워크를 거치는 방법이 연구되면 추천의 성능 또한 높아질 것으로 판단된다. 또한 시청의 패턴이 비교적 일정한 시청자는 결과의 성공률이 매우 높았지만, 시청의 패턴이 일정하지 않은 즉, 선호도를 일정하게 정의하기 어려운 시청자의 경우는 추천의 성공률이 비교적 낮았다. 이런 문제는 지상파 채널의 경우 채널의 특성이 차별화 되지 않고 여러 장르와 세부 장르가 공존하기 때문에 특성에 따른 추천이 어려웠기 때문으로 판단된다. 추천 연구를 케이블 TV와 디지털 TV방송 채널까지 확대하여 연구한다면 이런 문제를 해결할 수 있을 것이다. 마지막으로 시간 확장의 성과를 판단하기 위해 기간에 따라 가중치를 적용한 실험과 모든 기간의 가중치를 1로 동일하게 적용한 실험을 비교 했을 때, 가중치를 적용한 실험이 더 높은 추천 성공률을 보였다. 이는 최근의 시청 기록이 과거의 시청 기록보다 사용자가 시청할 방송 선택에 더 큰 영향을 미친다는 전제가 적절하게 반영된 것으로 보인다.

[참 고 문 헌]

[1] Yolanda Blanco-Fernandez, "A Multi-Agent Open Architecture for a TV Recommender System A case Study using a Bayesian Strategy",Proceedings of the IEEE , Sixth International Symposium on Multimedia Software Engineering, 2004

[2] 김용, 멀티미디어 콘텐츠를 위한 이용빈도 기반 하이브리드 추천시스템에 관한 연구, 연세대학교 문헌정보학과 박사학위 논문, 2006

[3] 고민정, 경매 시스템에서 시계열 분석에 기반한 낙찰 예정가 추천 방법, 동국대학교 컴퓨터공학과 박사학위 논문, 2005

[4] 최준혁 외, 군집분석과 베이시안 학습을 이용한 웹 도서 동적 추천 시스템, 한국 퍼지 및 지능 시스템 학회, Vol.12 No.5, 2002

[5] 김달호, R과 WINBUGS를 이용한 베이지안 통계학, 자유아카데미, 2005