개인화된 방송 컨텐츠 추천을 위한 가중치 적용 Markov 모델

(Weighted Markov Model for Recommending Personalized Broadcasting Contents)

박 성 준 † 홍 종 규 †† 강 상 길 ††† 김 영 국 ††††
(Sung-joon Park) (Jong-kyu Hong) (Sang-gil Kang) (Young-Kuk Kim)

요 약 본 논문에서는 시간에 따라 다양한 컨텐츠를 제공하는 방송 환경에서 고객의 최근 시청 정보를 이용하여 바로 다음에 고객이 시청하기를 선호하는 컨텐츠를 추천하기 위한 방법으로 가중치 적용 Markov 모델을 제안한다. 일반적으로 TV 시청자들은 최근에 시청한 자신이 선호하는 컨텐츠를 다시 시청하는 성향이 있다. 본 논문에서 제안하는 가중치 적용 Markov 모델은 TV 시청자들의 이와 같은 성향을 고려하여 고객이 연속적으로 시청한 정도에 따라 컨텐츠 선호도 전이 행렬에 가중치를 적용한다. 제안된 모델의 실험을 위해 고객으로부터 수집된 TV 시청 정보를 이용하여 고객의 선호 장르를 추천하는데 제안 모델을 적용하였다. 실험 결과 제안된 방법이 기존 방법에 비해 추천의 정확도가 향상되었음을 보인다.

키워드 : 방송 컨텐츠, 개인화, 추천, Markov 모델

Abstract In this paper, we propose the weighted Markov model for recommending the users' prefered contents in the environment with considering the users' transition of their content consumption mind according to the kind of contents providing in time. In general, TV viewers have an intention to consume again the preferred contents consumed in recent by them. In order to take into the consideration, we modify the preference transition matrix by providing weights to the consecutively consumed contents for recommending the users' preferred contents. We applied the proposed model to the recommendation of TV viewer's genre preference. The experimental result shows that our method is more efficient than the typical methods.

Key words : Broadcasting Contents, Personalization, Recommendation, Markov model

1. 서 론

최근 디지털 TV 채널, 인터넷, 모바일 등에 대한 컨텐츠 또는 정보의 양이 거대하게 증가하면서 고객은 때때로 자신이 원하는 컨텐츠를 찾는데 어려움을 겪게 되며, 많은 시간을 소비하게 된다. 예를 들어, 디지털 멀티미디어 방송(DMB: Digital Multimedia Broadcasting)은 서비스되는 채널 수가 엄청나게 많기 때문에 시청자들은 때로 자신이 선호하는 채널을 선택하는데 많은 시간을 소비한다. 심지어는 자신이 선호하는 컨텐츠를 찾는 동안 이미 방송이 끝날 수도 있다. 이와 같은 문제를 해결하기 위한 방법으로 고객이 선호하는 컨텐츠를 미리 예측하여 추천함으로써 어느 정도 보다 편리한 생활을 고객에게 제공할 수 있다[1,2].

불특정 다수에게 TV 프로그램을 제공하는 기존의 방송 서비스 개념을 시청자가 선호하는 TV 프로그램 중심으로 시청할 수 있도록 하는 개인화된 방송 서비스가 DMB 방송 시대의 주요 요소가 되고 있다. 시간 변화에 따라서 고객이 선호하는 컨텐츠를 추천해 주는 개인화 방송은 고객이 선호하는 프로그램을 찾는데 걸리는 시간을 줄여줄 수 있으며, 선호 프로그램을 찾는 동안 이미 방송이 진행되어 원하는 방송을 놓쳐버리는 경우를 줄여줄 수 있다.

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성 지원 사업의 연구결과로 수행되었음(IITA-2005-C1090-0502-0016)
† 정 회 원 : 공주영상대학 모바일게임과 교수
sjpark@kcac.ac.kr
†† 학생회원 : 충남대학교 컴퓨터공학과
jkhong@cnu.ac.kr
††† 정 회 원 : 인하대학교 컴퓨터공학부
sgkang@inha.ac.kr
†††† 정 회 원 : 충남대학교 컴퓨터공학과 교수
ykim@cnu.ac.kr
논문접수 : 2006년 2월 6일
심사완료 : 2006년 8월 24일

본 논문에서는 이와 같이 시간에 따라 다양한 종류의 컨텐츠를 제공하는 방송 환경에서 고객의 최근 시청 정보를 이용하여 바로 다음에 고객이 시청하기를 선호하는 컨텐츠를 추천하기 위한 방법으로 가중치 적용 Markov 모델을 제안한다. 가중치 적용 Markov 모델은 고객이 최근에 시청한 컨텐츠 전이 행렬에 고객이 시간 변화에 따라 연속적으로 시청한 컨텐츠에 가중치를 고려한다. 실험 결과 제안된 방법이 기존 방법에 비해 추천의 정확도가 향상되었음을 보인다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련 연구에 관하여 서술하고, 제3장에서는 전형적인 Markov 모델에 대하여 살펴본다. 제4장에서는 시간 변화에 따른 고객의 선호도 전이 행렬에 고객 행위의 연속성을 반영한 가중치 적용 Markov 모델에 대해 기술하고, 제5장에서는 본 논문에서 제안하는 가중치 적용 Markov 모델의 타당성 검증을 위한 실험 결과를 기술한다. 그리고 마지막으로 제6장에서 결론을 맺는다.

2. 관련 연구

2.1 개인화(Personalization)

개인화는 고객이 원하거나 필요로 하는 정보를 제공하여, 이를 찾는데 걸리는 시간과 비용을 절약해 주고, 손쉽게 접근하도록 고객 선호도에 따라 동적으로 제공하는 것이다. 누구나 똑 같이 정적인 웹 페이지만을 제공하던 방식에서 벗어나 개인의 특성에 따라 또는 자신과 유사한 고객으로부터 동적인 웹 페이지를 통해 고객이 원하거나 필요로 하는 정보를 제공한다. 개인화는 웹 사이트를 중심으로 많은 연구가 되어왔으며, 최근에는 모바일 환경에서 모바일 장치가 가지는 한계를 극복하고, 모바일의 특성을 살리기 위한 방법으로 개인화에 대한 중요성이 더욱 강조되고 있다. 모바일 장치는 화면이 작고 대역폭이 낮기 때문에 한 화면에 많은 내용을 보여줄 수 없으며, 다량의 정보를 제공하는데 시간이 오래 걸린다. 따라서 개인화를 통해 불필요한 정보는 제외하고 고객 자신에게 필요한 정보만을 제공함으로써 이를 극복할 수 있다.

개인화 추천 시스템은 크게 여과(Filtering)방식과 추론(Inference) 방식으로 구분할 수 있다. 여과 방식은 고객 성향에 따른 추천 정보의 여과로써, 여과 방법에 따라 규칙 기반 필터링(Rule-Based Filtering)과 내용 기반 필터링(Content-Based Filtering) 그리고 협업 필터링(Collaborative Filtering)으로 나눌 수 있다[3]. 추론 방식은 고객의 행위에 따른 과거 정보를 기반으로 고객이 선호하는 컨텐츠를 추론하는 방식이다.

2.1.1 규칙 기반 필터링

규칙 기반 필터링(Rule-Based Filtering)은 사용자의 과거 행위나 개인 신상, 관심 분야, 선호도 등에 대해 사용자로부터 얻은 명시적 프로파일 정보를 이용하여 미리 생성된 규칙을 가지고 적용하는 기법으로 if~then 규칙[4], 결정 트리 구조[5]가 가장 많이 사용된다. Kim et al.[6]은 트리 추론 방법[5]을 이용하여 인터넷 상점 초기 화면에 개인화 추천을 제공하기 위한 마케팅 규칙 추출 기법을 제안하였다. 대표적인 규칙 기반 기술 중 하나로서, Aggrawall et al.[7, 8]은 연관 규칙 마이닝 알고리즘[9]을 이용하여 추정 빈도 분포로부터 자주 발생하는 아이템 집합을 찾아내는 방법을 제안하였다. 이와 같은 규칙은 추천 엔진이 간단하다는 장점이 있으나, 이미 구조화된 규칙에 제한적이며, 규칙을 생성하는 전문가에게 의존적이므로 규칙 생성이 체계적이지 못하다. 그리고 이미 정의된 규칙을 변경하기 위해서는 임의로 조작해야 하는 번거로움이 있다.

2.1.2 내용 기반 필터링

내용 기반 필터링(Content-Based Filtering)은 과거에 목표 고객이 선호했던 아이템과 가장 유사한 아이템을 찾아 추천하는 방식이다. 고객으로부터 이미 평가된 아이템과 이와 관련된 내용을 분석하여 고객이 이전에 선호한 아이템과 비슷한 특성을 갖는 항목을 선호할 가능성이 높다고 보고 선호도가 표시된 항목들의 속성 정보를 이용하여 추천하는 기술이다[10].

내용 기반 필터링의 장점은 아이템 자체를 모델링하는 기법이기 때문에 단순하다. 또한 전반적인 고객 그룹이 이질적인 평가를 보이는 상품 그룹의 추천에 정확도가 높다. 단점으로는 첫째, 각 아이템에 대한 특성을 추출하고 이를 기반으로 추천 대상을 정하게 되므로 효과적으로 이루어지기 어렵다. 둘째는 고객이 이전에 좋게 평가한 아이템과 유사한 아이템을 추천하므로 추천 결과가 특정 부분으로 치우치게 될 수 있다. 셋째는 고객이 아이템에 대한 인식과 선호도를 공식화하는데 문제가 있다. 어느 한 고객이 특정 아이템을 좋아하는지, 싫어하는지 또는 다른 것에 비해 특정 아이템을 왜 선호하는지 그 이유를 가상으로 공식화하는 것이 어렵다.

2.1.3 협업 필터링

협업 필터링은 오늘날 대부분의 성공적인 추천 시스템에서 가장 많이 쓰이는 대표적인 방법으로 소비 성향이 비슷한 취향을 가진 고객 또는 아이템을 그룹으로 묶어서 목표 고객(Target User)과 취향이 유사한 고객을 찾아서 그 고객이 좋아하는 아이템을 추천하는 방법이다. 협업 필터링은 기술의 성숙도로 net-news[11], e-commerce[12-14], 디지털 도서관[15], 디지털 TV [16-18]와 같은 여러 가지 다양한 선호도를 예측하는데 매력적으로 이용되어 왔다. 일반적으로 규칙 기반 필터링 및 협업 필터링 기법은 많은 고객들의 소비 행위를

수집하기 때문에 많은 노력, 시간, 비용이 필요하다. 협업 필터링 방식은 사용자 기반 협업 필터링 방식과 아이템 기반 협력적 필터링 방식으로 나눌 수 있다.

사용자 기반 협업 필터링은 유사한 성향을 가지는 고객들을 그룹으로 묶어서 목표 고객의 선호도와 가장 유사한 선호도를 가진 고객 그룹에 의해 선정된 아이템 또는 가장 높은 점수를 가지는 아이템을 추천하는 방식이다. 즉, 아이템 A를 선택한 고객들의 그룹은 아이템 B와 C를 선호한다고 가정하자. 그러면 아이템 A를 선택한 목표 고객(Target User)은 아이템 A를 선택한 고객 그룹을 찾아 그 그룹이 선호하는 아이템 B와 C를 추천해 주는 방식이다.

아이템 기반 협업 필터링은 고객이 선호도를 입력한 기존 상품들과 예측하고자 하는 상품의 상관 관계를 계산하여 선호도를 예측한다. 이 방법에서는 상품들 간의 유사도를 계산하기 위하여 두 상품에 대해 선호도를 입력한 고객들의 정보를 이용한다. 예를 들어, 아이템 A, B, C, D에 대해 평가한 고객들의 선호도를 바탕으로 각 아이템들간의 유사도를 계산한다. 목표 고객이 선호하는 아이템이 A라고 하자. 그러면 아이템간의 유사도 테이블에서 아이템 A와 가장 유사한 아이템을 찾는다. 즉, 목표 고객이 선택한 아이템 A 에 대한 선호도 평가와 가장 유사한 아이템을 추천하는 방식이다.

협업 필터링은 아이템의 질적인 면과 선호도에 기반하여 아이템을 여과하는 능력이 있기 때문에 사용자가 원하는 내용이 아이템에는 포함되어 있지 않아도 유사한 선호도를 가지는 사용자들에게 좋은 평가를 얻었다면 그 아이템을 추천할 수 있다. 즉 유사 사용자의 평가값을 이용하여 현 사용자의 평가 값을 예측하는 방법이므로 내용기반 방식에 비하여 우수한 상품을 추천할 수 있다. 또한 내용 기반 필터링에 비해 협업 필터링은 알고리즘이 간단하고 고객이 관심을 가지는 새로운 아이템도 추천할 수 있다. 그러나 새로운 아이템에 대한 고객의 평가 정보가 없는 경우는 상품의 선호도 값을 예측할 수 없다는 단점을 가지고 있다. 즉 희박성 문제와 초기 사용자 문제를 가진다.

2.1.4 추론 방식

고객의 과거 이용 정보를 기반으로 고객의 소비 행위를 학습하고 가까운 미래에 고객이 선호할 컨텐츠를 추론하는 방식이다. 예를 들어, 고객의 클릭 정보를 모니터링 한다고 가정하자. 적어도 10번 중 7번은 컴퓨터 서적을 클릭한다면, 이 고객은 컴퓨터 서적에 관심을 가지고 있다고 추론함으로써 컴퓨터 서적과 관련된 정보를 더 많이 제공하도록 한다. 추론 방식은 규칙 기반 추천 시스템과 유사 하지만, 고객의 최근 정보를 계속 모니터링 해서 학습을 통해 실시간으로 고객의 선호도를 추적할 수 있다. 추론 방식은 묵시적 데이타만을 이용해서 고객의 선호도를 예측할 수 있다. 추론 방식으로는 결정트리, 베이지안 네트워크, 그리고 Markov 모델 등이 있다.

규칙 기반 필터링과 협업 필터링은 많은 고객의 소비 행위 정보가 필요하므로 시간, 비용, 노력이 많이 요구되는 반면, 추론 방식은 고객 개인의 소비 행위에 대한 정보를 기반으로 고객의 소비 행위를 예측하기 때문에 상대적으로 정보의 양이 적고, 비용이 적게 들어간다. Brown et al. [19]은 베이지안 네트워크[20]를 특정 고객 모델링에 동적으로 적응하도록 하는 방법론을 제안하였다. Lin et al. [21]는 입력하는 문장, 멀티미디어 파일을 삽입하거나 전송하는 고객의 상호작용으로부터 얻어진 특징이나 행위를 기반으로 통계 예측 모델을 체계화 하였다. 그 외 몇몇 연구들[22,23]은 대응하는 베이지안 네트워크로부터 생성된 표본을 이용하여 훈련된 이진, 2계층, 노이지-OR 네트워크[24]로부터 의료 진단을 위해서 사후 확률 분포를 추론한다. 근사값 분포는 실제 사후 값에 적합도 측정값을 최대화하는 결정 알고리즘에 의해 얻어진다.

지금까지 언급된 연구로부터 고객이 선호하는 컨텐츠를 예측하거나 추천하는 대부분의 방법들은 고객에 의해 직접 시스템에 입력하는 명시적 고객 프로파일 또는 고객의 과거 소비 행위를 이용한다. 현재의 선호도에서 다음 상태로의 변환 정보는 컨텐츠에 대한 고객의 향후 선호도를 예측하는데 주요 단서가 될 수 있는 연구에는 이용되지 않았다. 이와 같은 문제를 해결하기 위하여 많은 학자[25-27]들이 Markov 모델을 이용하였다. 전형적인 Markov 모델은 시간에 따른 고객의 과거 통계적 선호도 전이 정보만을 이용하여 가까운 미래의 고객 선호도를 예측한다. 일반적으로 고객들은 자신이 최근에 이용한 선호하는 컨텐츠를 다시 이용하는 경향이 있다. 이런 점을 고려하여 본 논문에서는 고객의 선호 컨텐츠를 추천하는데 최근에 연속적으로 소비된 컨텐츠에 연속성의 정도에 따라 가중치를 부여함으로써 전형적인 Markov 모델을 확장한다.

2.2 방송 컨텐츠 개인화 방식

방송 컨텐츠 정보를 제공하기 위한 개인화 시스템은 크게 두 가지로 분류할 수 있다. 첫째는 디지털 방송국과 같이 멀티미디어 컨텐츠를 제공하는 서버 쪽에서 개인화를 제공하는 방식이다. 다른 하나는 모바일 장치와 같이 멀티미디어 컨텐츠를 제공 받는 클라이언트 쪽에서 개인화를 제공하는 방식이다. 이와 같은 개인화 서비스를 제공하기 위한 기술은 크게 두 가지가 있다. 하나는 협업 필터링[28]이고, 다른 하나는 추론 방식[29]이다. 협업 필터링은 고객과 비슷한 소비 성향을 가지는 고객 그룹을 기반으로 고객이 선호로 하는 컨텐츠를 결

정한다. 추론 방식은 고객의 과거 이용 정보를 기반으로 고객이 선호로 하는 컨텐츠를 추론한다. 서버 쪽에서 제공되는 개인화는 일반적으로 클라이언트 쪽에서 제공되는 개인화와 달리 협업 필터링에 의해서 이루어 진다. 서버 쪽 개인화 시스템의 예로는 Cotter et al.[30]이 TV 프로그램에 대한 고객의 명시적 선호 점수를 기반으로 하는 웹 기반 개인화 전자 프로그램 가이드인 PTV[31]를 소개한다. Lee et al.[32]는 서버 쪽에서 자동화된 채널 추천을 위한 고객 중심의 리모콘 시스템을 데모한다. Konstan et al.[33]은 방대한 양의 문서들 중에서 선호 문서를 찾아 주는 시스템인 GroupLens에 협업 필터링을 적용하였다.

멀티미디어 분야에서 개인화 연구의 대부분은 서버 쪽에 치우쳐 있다. 이러한 경우 서버 쪽에서 취급되는 정보의 양은 고객의 수가 점점 늘어나면서 과다해 진다. 이와 같은 문제를 해결하기 위하여 클라이언트 중심의 개인화가 연구되어 왔다[2,34]. 클라이언트 쪽에서는 추론하는데 고객 자신의 usage 정보만 이용하기 때문에, 클라이언트 쪽에서 고객이 선호하는 컨텐츠를 추론하기 위한 데이타의 양은 서버 쪽의 데이타 양에 비해 상대적으로 아주 작다. 클라이언트 쪽에서 실행되는 개인화의 가장 대표적인 예는 디지털 방송 시스템에서 개인화된 전자 프로그램 가이드(PEPG: Personalized Electronic Program Guide)이다. Lee et al.[35]는 고객의 과거 시청 기록으로부터 선호 장르에 대한 통계적 추청(TOP-N)을 이용한 전자 프로그램 가이드를 개발하였다. Setten[36]은 영화 장르에 대한 고객의 관심 정도를 예측하기 위한 장르 추론 방법을 제안하였다. 제안한 추론 방법은 장르 선형 함수를 이용하여 고객에 대한 각 TV 프로그램 장르의 상대적인 중요도를 학습하는 장르 최소자승법(GenreLMS)를 이용하였다. Kang et al.[2,37]는 TV 장르에 대한 현재 선호도와 과거 선호도간의 상호 정보를 이용하여 TV 장르를 추천하기 위한 알고리즘을 소개하였다.

3. Markov 모델

Markov 모델은 일반적으로 일련의 확률변수의 통계값을 예측하기 위해 이용된다. Markov 모델로부터 유도되는 기술은 시간에 따라 변하는 상태가 존재하며, 고객이 이미 수행한 연속적인 행위가 주어져 있으면, 고객이 다음에 행할 행위를 예측하기 위해 포괄적으로 이용된다. Markov 모델은 <행위(A), 상태(S), 전이(T)> 3개의 파라미터로 표현한다. 여기서, 행위(A)는 고객에 의해 수행될 수 있는 모든 가능한 행위들의 집합이다. 상태(S)는 Markov모델이 만든 모든 가능한 상태의 집합으로 다음과 같이 전이 행렬을 구할 수 있다.

T = |S| x |A|: 전이 행렬(Transition Matrix)

$$T = \begin{bmatrix} t_{11} & \cdot & \cdot & \cdot & t_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & t_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ t_{n1} & \cdot & \cdot & \cdot & t_{nn} \end{bmatrix}$$

여기서, 각각의 요소인 t_{ij}는 고객이 상태 i에 있을 때, 행위 j를 행할 빈도수이다.

Markov 모델의 전이 행렬은 다음 행위를 예측하는데 이용된 사전 행위의 수에 종속적이다. 가장 단순한 Markov 모델은 고객에 의해 수행된 마지막 행위만 보고 다음 행위를 예측하는 것으로 1차 Markov 모델이라 한다. 1차 Markov 모델의 전이 행렬에서 고객에 의해 수행될 수 있는 각각의 행위는 상태에 해당된다. 좀더 복잡한 모델은 고객에 의해서 수행된 마지막 2개의 행위를 보고 예측도를 계산하는 것이다. 이와 같은 것을 2차 Markov 모델이라 하며, 2차 Markov 모델의 상태는 연속적으로 수행될 수 모든 행위들의 쌍에 대응한다. 이와 같은 방법은 고객에 의해서 수행된 마지막 r개의 행위를 보고 예측도를 계산하는 r차 Markov모델로 일반화 할 수 있다. 이를 식으로 표현하면 다음과 같다.

일련의 확률변수 X를 $X(1), X(2), \cdots, X(t)$로 표시하자. 여기서 괄호 안의 숫자는 시간을 의미한다. 변수 X는 $X = \{x_1, x_2, \cdots, x_c\}$로 표시된 c개의 상태를 가진다. 과거 r개의 상태가 주어졌을 때, t+1 시점에서 $X = x_i$일 조건 확률은 식 (1)과 같이 표현할 수 있다.

$$P(X(t+1) = x_i \mid X(t), X(t-1), \cdots X(t-r+1)) \quad (1)$$

여기서, r은 Markov 모델의 차수를 의미한다.

따라서, 1차 Markov 모델은 r=1일 때를 의미하며, 식 (2)와 같이 표현할 수 있다.

$$P(X(t+1) = x_i \mid X(t)) \quad (2)$$

식 (1), 식 (2)에서 보는 것처럼 Markov 모델은 시간에 따라 상태가 변하는 고객의 소비 행위를 예측하는데 이용되어 왔다. Sarukkai[25]는 Markov 체인을 이용하여 확률적으로 링크를 예측하고 웹의 패스를 분석하는 방법을 제안하였다. Sarukkai[25]의 논문에서 고객의 웹 공간의 항해에 대한 표현으로 Markov 상태 전이 행렬을 사용하였다. Cadez et al.[26]는 Markov 모델을 이용하여 브라우징 세션을 분류하였다. [25,26]에서 이용된 것처럼, 전형적인 Markov 모델의 크기는 차수가 증가함에 따라 급속도로 증가한다. 이와 같은 단점을 해결하기 위한 방법으로 He et al.[27]은 웹 접속 예측을 위한 트리 형태의 Markov 모델을 구조화 하였다.

Markov 모델은 웹 사이트에서 사용자의 브라우징 행

위를 모델링하고 예측하는데 아주 적합함을 보였다. 일반적으로 이와 같은 문제를 위한 입력 데이타는 고객에 의해 접근된 일련의 웹 페이지이며, 목표는 모델에 이용될 수 있는 Markov모델을 만들어서 고객이 다음에 가장 접근하고 싶어하는 웹 페이지를 예측하는 것이다.

3.1 1차 Markov 모델

전통적인 Markov 모델은 식 (1)과 같다. 식 (1)에서 r=1일 때를 1차 Markov 모델이라 하며, 식 (2)와 같이 표현할 수 있다. 즉, 1차 Markov 모델은 그림 1과 같이 현재 시간 t시간에 소비하고 있는 컨텐츠의 종류를 알고 있을 때, 과거 컨텐츠 소비 성향 전이 확률 행렬을 이용하여 바로 다음 시간 t+1 시간에 어떤 종류의 컨텐츠를 소비하게 될지 선호 컨텐츠를 예측하기 위한 모델이다. 과거 컨텐츠 소비 성향 전이 확률 행렬은 19페이지의 표 1과 같이 나타낼 수 있다. 그림 1에서 원으로 표시된 것은 소비 성향의 상태를 의미한다.

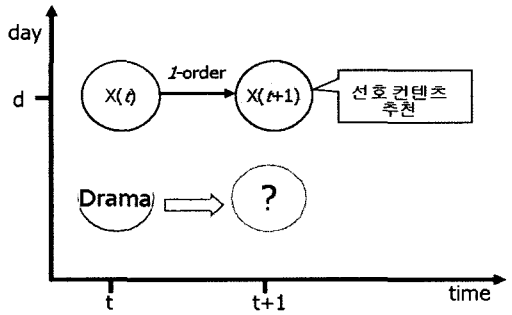

그림 1 1차 Markov 모델

그림 1에서 $X(t)$는 시간 구간 t에서 고객의 컨텐츠 소비 성향 변수를 의미한다. 변수는 [x_1, x_2, ⋯, x_i, ⋯, x_c]와 같이 c 개의 컨텐츠를 가질 수 있다. 여기서 x_i는 i^{th} 번째 컨텐츠를 의미한다. 즉, 변수 X={Drama, Sports, News, ⋯, Others}로 표현할 수 있다.

일정 기간 동안 고객의 컨텐츠 소비 기록으로부터 day d에 특정 시간 구간 t에서 다음 시간 구간 t+1 사이에 i^{th} 번째 컨텐츠 x_i에서 j^{th} 번째 컨텐츠 x_j로의 추정된 통계적 성향의 변화는 식 (3)과 같이 표현될 수 있다.

$$\hat{P}\left(X(t+1,d)=x_j \mid X(t,d)=x_i\right)=\frac{n_{ij}}{\sum_{j=1}^{c} n_{ij}} \qquad (3)$$

여기서, n_{ij}는 고객의 컨텐츠 소비 기록으로부터 day d에 시간 구간 t에서 x_i가 소비되었다는 사실이 주어졌을 때, 시간 구간 t+1에서 x_j를 소비한 빈도수를 의미한다. 그리고 j = 1, 2, ⋯, c이다. 식 (3)은 고객이 선호하는 컨텐츠를 예측하기 위한 1차 Markov 모델이다.

3.2 r차 Markov 모델

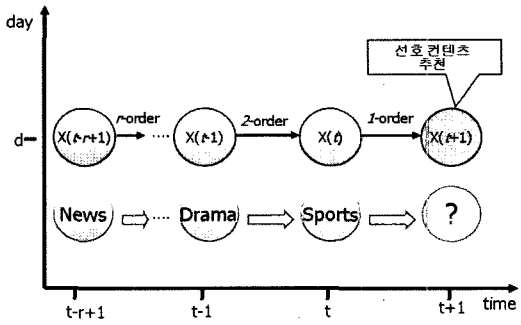

그림 2 r차 Markov 모델

1차 Markov 모델은 과거 연속적인 소비 성향의 변화를 r개까지 확장함으로써 r차 Markov 모델로 확장할 수 있다. 그림 2는 시간에 따라 연속적으로 발생하는 소비 성향 전이 상태를 r까지 확장했을 때의 Markov 모델을 가리키는 것으로 1차 Markov 모델을 일반화 한 것이 r차 Markov 모델이다. r차 Markov 모델은 식 (1)과 같이 과거 r개의 상태 전이가 주어졌을 때, t+1 시점에서 $X=x_i$일 조건 확률로 표현할 수 있으며, 식 (4)와 같이 표현할 수 있다.

$$\hat{P}\left(X(t+1,d)=x_j \mid X(t,d)=x_i, X(t-1,d)=x_k,\ldots,X(t-r+1,d)=x_m\right)=\frac{n_{x_j}}{\sum_{i=1}^{c} n_{x_i}} \qquad (4)$$

여기서, r은 Markov 모델의 차수를 의미하며, n_{x_i}는 고객의 컨텐츠 소비 기록으로부터 day d에 시간 구간 t-r+1과 t사이에 상태 전이의 주어진 패턴 $X(t,d)=x_i, X(t-1,d)=x_k,\ldots,X(t-r+1,d)=x_m$에 대해서 day d에 시간 구간 t+1에서 소비하는 컨텐츠 x_i의 빈도수이다. 각 패턴의 빈도수는 고객의 컨텐츠 소비 기록으로부터 사전에 정해진 간격(1일, 1주일, 등)으로 이동하면서 각 패턴의 발생 빈도 수를 계산함으로써 얻어질 수 있다. 일반적으로 높은 차수의 Markov 모델이 낮은 차수의 Markov모델보다 더 세부적이기 때문에 더 높은 차수의 Markov 모델이 낮은 차수의 Markov모델보다 예측에 대한 정확도가 더 높다. 그러나 이와는 대조적으로 발생 가능한 패턴의 조합이 증가함으로써 소비 기록이 충분하지 않은 경우 low coverage가 발생할 수 있으며, 발생 가능한 상태의 수가 지수적으로 증가하는 경향이 있다. 그러므로 Markov 모델에서 차수의 선정은 고객의 컨텐츠 소비 기록의 크기에 종속적이다[38,39].

식 (3)과 (4)는 단지 하나의 주어진 시간 구간 t에서 하나의 컨텐츠만 소비될 때, 시간 구간 t+1에서 컨텐츠 x_i의 통계적 선호도를 추정하는데 적합하다. 그러나 시간 구간의 크기에 따라 여러 개의 컨텐츠가 소비될 수 있다. 예를 들어, 시간 구간의 크기가 1시간인 경우 1시간 동안 드라마도 시청하고 스포츠도 시청할 수 있다. 만약 시간 구간 t에서 소비된 컨텐츠의 집합을 c′이라 하면, c'= {드라마, 스포츠}가 된다. 따라서, 식 (3)은 식 (5)와 같이 다시 작성할 수 있다.

$$\hat{P}\left(X(t+1,d)=x_j \mid X(t,d)=c'\right)=\frac{\sum_{i\in c'} n_{ij}}{\sum_{j=1}^{c}\sum_{i\in c'} n_{ij}} \qquad (5)$$

같은 방법으로 r차 Markov 모델의 경우, 식 (4)로부터 다음과 같이 수정할 수 있다.

$$\hat{P}\left(X(t+1,d)=x_j \mid \{X(t,d), X(t-1,d),\ldots, X(t-r+1,d)\}=C'\right)=\frac{\sum_{C'} n_{x_j}}{\sum_{j=1}^{c}\sum_{C'} n_{x_j}} \qquad (6)$$

여기서, C'는 고객의 컨텐츠 소비 기록으로부터 day d 에서 시간 구간 t-r+1와 t 사이의 상태 전이 패턴의 집합 $[X(t,d),\ X(t-1,d),\ \ldots, X(t-r+1,d)]$이다.

4. 가중치 적용 Markov 모델

기존 Markov 모델은 특정 날 *day d*에서의 시간에 따른 컨텐츠 소비 행위만을 고려한 것이다. 그러나 방송 컨텐츠에 대한 고객의 소비 성향은 시간 변화에 따라 다양하게 전이되지만, *day d*로부터 k 만큼 이전 day 즉, d-1, d-2, ⋯, d-k 에 방송된 컨텐츠가 고객의 소비 성향에 많은 연관이 있을 수 있다. 어느 한 시청자가 day d-1의 시간 구간 t에서 어느 한 TV 프로그램을 보았으며, 이 프로그램은 이틀간 연속해서 방영된다고 가정하자. 그러면, 그 시청자는 어제 본 그 TV 프로그램이 자신이 선호하는 프로그램이었다면, *day d*에도 또 시청하려 할 것이다. 예를 들어, 어학 공부에 관심이 있는 고객은 매일 같은 시간에 어학 강좌를 시청하려 할 것이다. 이와 같이 요일에 따라 연속해서 같은 시간에 동일 컨텐츠를 시청하는 경우 이 컨텐츠를 선호하는 것으로 가정하고 고객의 선호도 예측에 이를 반영하기 위한 방법이 가중치 적용 Markov 모델이다.

그림 3은 고객의 방송 컨텐츠 선호도를 예측하기 위한 가중치 적용 1차 Markov 모델의 예이다. 가중치 적용 1차 Markov 모델은 전형적인 1차 Markov 모델에 그림 3과 같이 2일 이상 같은 시간에 시청한 동일 컨텐츠에 대해서만 가중치($W^j_{d-1,d}$)를 적용한다. 식 (5)은 특정 날 *day d*에서의 컨텐츠 소비 행위만을 고려한 1차 Markov 모델이다. 따라서, 하루 전날 *day d*-1 동일 시간에 소비한 행위를 적용한 가중치 적용 1차 Markov 모델은 식 (7)과 같이 구할 수 있다.

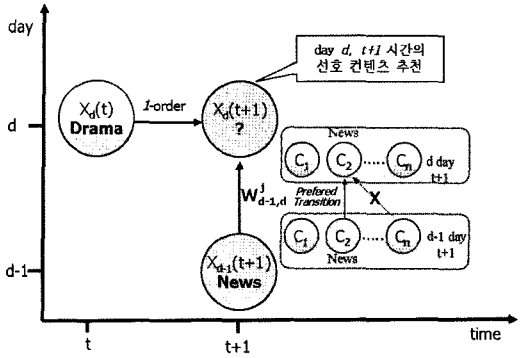

그림 3 가중치 적용 1차 Markov 모델의 예

$$\hat{P}_W\left(X(t+1,d)=x_j \mid X(t,d)=c'\right)=\frac{\sum_{i\in c'} w^j_{d-1,d}(t+1)n_{ij}}{\sum_{j=1}^{c}\sum_{i\in c'} w^j_{d-1,d}(t+1)n_{ij}} \qquad (7)$$

여기서,

$$\left\{ w^j_{d-1,d}(t+1)=\begin{cases} \rho(1+\eta^j_{d,d-1}(t+1)), & d-1\text{일}, t+1\text{시간에 } x_j \text{ 를 소비한 경우} \\ 1, & \text{그외의 경우} \end{cases} \right\} \qquad (8)$$

여기서, ρ는 일반적으로 경험적 실험에 의해서 최적의 값을 선정할 수 있는 scaling factor이다.

$$\eta^j_{d-1,d}(t+1)=\frac{n_{x_j,d-1,d}(t+1)}{n_{x_j,d-1}(t+1)} \qquad (9)$$

여기서, $n_{x_j,d-1}(t+1)$와 $n_{x_j,d-1,d}(t+1)$는 고객의 컨텐츠 소비 기록으로부터 d-1 day에 시간 구간 t+1에서 소비된 x_j의 빈도수와 d-1과 d day에 시간 구간 t+1에서 소비된 x_j의 빈도수이다.

그림 4는 d day부터 r day 전까지 t+1 시간에 소비된 컨텐츠를 고려하여 가중치를 부여하였을 때의 r 차 Markov 모델이다. 지금까지는 d day부터 하루 전날인 d-1에 소비된 고객 컨텐츠의 상관관계가 컨텐츠의 통계적 선호도를 추정하는데 이용되었다. d day부터 하루 이상의 전날 즉, d-k, d-k+1, ⋯, d-1을 고려함으로써 고객 행위에 대한 정보는 더욱더 풍부해 진다. 앞에서 언급한 것과 같이 컨텐츠가 연속해서 소비되면 될수록, 고객이 가까운 미래에 소비할 수 있는 가능성이 점점 더 높아진다. 이와 같이 기존의 r 차 Markov 모델에

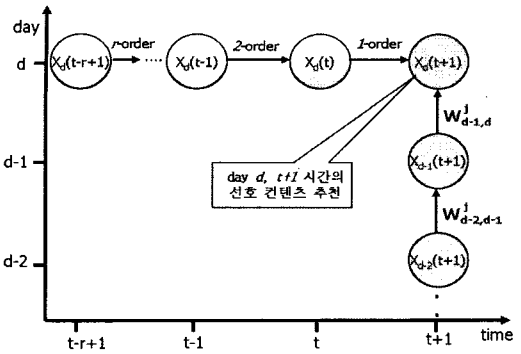

그림 4 가중치 적용 r 차 Markov 모델

day를 고려하여 같은 시간에 동일한 컨텐츠를 소비하는 경우 가중치를 적용한 r 차 Markov 모델이라 한다.

가중치 적용 r 차 Markov 모델은 그림 4와 같이 연속해서 몇 일간 시간 구간 t+1에서 컨텐츠를 소비하면 할수록, 그 컨텐츠에 대한 소비 빈도수에 더 많은 가중치가 제공된다. 선호도를 추정하는데 이와 같은 상황을 고려한 가중치 적용 r 차 Markov 모델은 r 차 Markov 모델인 식 (7)로부터 식 (10)과 같이 구할 수 있다.

$$\hat{P}_W\left(X(t+1,d)=x_j \mid X(t,d)=c'\right)=\frac{\sum_{i\in c'} W^j_{d-k,d}(t+1)n_{ij}}{\sum_{j=1}^{c}\sum_{i\in c'} W^j_{d-k,d}(t+1)n_{ij}} \quad (10)$$

여기서,

$$W^j_{d-k,d}(t+1)=\prod_{i=1}^{k} w^j_{d-i,d-i+1}(t+1) \quad (11)$$

여기서, $W^j_{d-k,d}(t+1)$는 컨텐츠를 두 번 연속해서 소비하는 쌍들의 곱으로 표현한다. 왜냐하면, k의 값이 커지는 경우, k 번 연속해서 소비하는 경우는 드물기 때문이다. 이런 경우 가중치를 적용할 수 없는 경우가 발생할 수 있다. 이와 같은 사항을 피하기 위해 d-k와 d day 사이에 2번 연속해서 사용하는 경우 가중치를 부여하는 것으로 하였다. 따라서 두 번 연속해서 소비하는 횟수에 의해 가중치가 적용된다.

가중치 적용 r차 Markov 모델에 의해 컨텐츠의 통계적 선호도를 추정하기 위한 수학적 표현은 주어진 시간 구간 t에서 여러 개의 컨텐츠를 소비하는 경우를 가정하면, 식 (5) 대신에 식 (6)을 이용하여 r 차 Markov 모델에 직접 적용할 수 있다. 그 결과 식 (12)와 같은 공식을 구할 수 있다.

$$\hat{P}\left(X(t+1,d)=x_j \mid \{X(t,d),X(t-1,d),\ldots,X(t-r+1,d)\}=C'\right)=\frac{\sum_{C'} w^j_{d-1,d}(t+1)n_{x_j}}{\sum_{j=1}^{c}\sum_{C'} w^j_{d-1,d}(t+1)n_{x_j}} \quad (12)$$

전형적인 Markov 모델과 가중치 적용 Markov모델에 대한 알고리즘을 분석하기 위해 4장에서 실험 데이타로 이용하기 위해 수집된 TV 시청 데이타 중 한 명의 시청자 데이타를 이용하여 가중치 적용 Markov 모델을 분석한다. 수집된 TV 시청 데이타 중 한 명의 시청자(Person A라 부르자)에 대한 TV 프로그램 시청 데이타를 이용하여 전형적인 Markov모델과 가중치 적용 Markov모델을 이용하여 분석해 보면 가중치 적용 1차 Markov 모델의 예측 성능이 전형적인 1차 Markov 모델에 비해 어떻게 향상될 수 있는지 확인해 볼 수 있다. 분석에 이용된 Person A는 30대 여성이다. 성능의 정확도(accuracy)는 다음과 같은 절차에 의해 구할 수 있다.

첫째, 고객으로부터 수집된 고객의 과거 TV 시청 데이타를 이용하여 8개 장르에 대한 선호도 전이 행렬을 생성한다.

둘째, 첫 번째 절차에서 계산된 값으로부터 최상위 값을 가지는 2개의 선호 장르를 추천한다. 여기서 2개를 추천하는 이유는 수집 데이타로부터 1시간에 시청하는 평균 TV 프로그램 수가 약 2개이기 때문이다.

셋째, 테스트 데이타에서 d day, 시간 구간 t+1에 시청한 실제 장르와 추천된 장르를 비교하여 일치하는 장르의 비율로 정확도를 계산한다.

다음은 전형적인 Markov모델을 이용하여 고객의 선호 장르를 추천하기 위한 과정이다. 표 1은 훈련데이타(2002년 12월 7일부터 2003년 5월 6일까지)로부터 d day(수요일)에 시간 구간 t(오후 7~8시)와 t+1(오후 8시~9시) 사이에 8개 장르에 대한 선호도 전이에 대한 확률 값을 계산하는 과정을 보이고 있다. 표 2는 테스트 데이타로부터 2003년 5월 7일 d day(수요일) 시간 구간 t (오후 7~8시)에 시청한 TV 장르를 나타내는 것으로 시청한 장르는 "1"로 표시하였다. 표 3은 테스트 데이타로부터 2003년 5월 7일 d day(수요일) 시간 구간 t+1 (오후 8~9시)에 시청한 TV 장르를 나타내는 것으로 추천 결과에 대한 정확도를 계산하기 위한 정보이다.

각 장르에 대한 선호도 예측 확률은 식 (5)를 이용하여 구할 수 있다. 실험데이타에서 시간 구간 t에서 시청한 장르의 집합은 표 2에서 보는 것처럼 C'={뉴스, 정보}이다. 따라서 표 1에서 밝게 표시된 두 개의 행에 속하는 장르에 대한 선호도 전이 데이타만을 이용하여 계산한다. 계산된 선호도 전이 값으로부터 뉴스와 정보가 추천될 수 있다. 추천 결과에 대한 정확도 계산을 위해 표 3을 확인해 보면 테스트 데이타로부터 d day에 시간 구간 t+1에 실제 시청한 장르는 뉴스와 드라마&영화이다. 따라서 전형적인 Markov모델의 정확도는 50%가 된다. 다음은 가중치 적용 Markov 모델을 이용하여 방

표 1 훈련 데이타로부터 1차 Markov 모델을 이용한 선호도 계산

		2002년 12월 7일부터 2003년 5월 6일까지 d day(수요일) 시간 구간 t+1 (오후 8~9시)에 시청한 TV 장르의 빈도수							
		뉴스	오락	드라마&영화	정보	스포츠	교육	어린이	기타
2002년 12월 7일부터 2003년 5월 6일까지 d day(수요일) 시간 구간 t (오후 7~8시)에 시청한 TV 장르의 빈도수	뉴스	3	0	2	3	1	0	0	0
	오락	4	0	3	3	1	0	0	0
	드라마&영화	2	0	3	2	0	0	0	0
	정보	6	0	4	5	0	0	0	0
	스포츠	0	0	1	1	0	0	0	0
	교육	0	0	0	0	0	1	0	0
	어린이	0	0	0	0	0	0	2	0
	기타	0	0	0	0	0	0	0	0
$\sum_{C'} n_{i_j}$ C'={뉴스, 정보}		9	0	6	8	1	0	0	0
$\frac{\sum_{i \in c'} n_{ij}}{\sum_{j=1}^{c} \sum_{i \in c'} n_{ij}}$		0.375	0.000	0.250	0.333	0.042	0.000	0.000	0.000

표 2 테스트 데이타로부터 t 시간에 시청한 TV 장르

2003년 5월 7일 수요일 시간 구간 t (오후 7~8시) 시청 정보							
뉴스	오락	드라마&영화	정보	스포츠	교육	어린이	기타
1	0	0	1	0	0	0	0

표 3 테스트 데이타로부터 t+1시간에 시청한 TV 장르

2003년 5월 7일 수요일 시간 구간 t+1(오후 8~9시) 시청 정보							
뉴스	오락	드라마&영화	정보	스포츠	교육	어린이	기타
1	0	1	0	0	0	0	0

표 4 훈련 데이타로부터 가중치 적용 Markov 모델을 이용한 선호도 계산

		2002년 12월 7일부터 2003년 5월 6일까지 d day(수요일) 시간 구간 t+1 (오후 8~9시)에 시청한 TV 장르의 빈도수								
		뉴스	오락	드라마&영화	정보	스포츠	교육	어린이	기타	
2002년 12월 7일부터 2003년 5월 6일까지 d-1 day(화요일) 시간 구간 t (오후 7~8시)에 시청한 TV 장르의 빈도수	뉴스	9	0	2	6	4	0	0	0	21
	오락	0	0	0	0	0	0	0	0	0
	드라마&영화	2	0	11	4	2	0	0	0	19
	정보	3	1	0	3	2	0	0	0	9
	스포츠	4	0	1	2	1	0	0	0	8
	교육	0	0	0	0	0	0	0	0	0
	어린이	0	0	0	0	0	0	0	0	0
	기타	0	0	0	0	0	0	0	0	0
$w_{d-1,d}^{j}(t+1)$		1.429	1.000	1.611	1.333	1.125	1.000	1.000	1.000	
$\sum_{i \in C'} n_{i_j}$ C'={ 뉴스, 정보 }		9	0	6	8	1	0	0	0	
$\sum_{i \in c'} w_{d-1,d}^{j}(t+1) n_{ij}$		12.857	0.000	9.667	8.000	1.000	0.000	0.000	0.000	31.524
$\frac{\sum_{i \in c'} w_{d-1,d}^{j} n_{ij}}{\sum_{j=1}^{c} \sum_{i \in c'} w_{d-1,d}^{j} n_{ij}}$		0.408	0.000	0.307	0.254	1.000	0.000	0.000	0.000	

표 5 테스트 데이타로부터 d-1, t+1시간에 시청한 TV 장르

2003년 5월 6일 d-1 day(화요일) 시간 구간 t+1(오후 8~9시) 시청 정보							
뉴스	오락	드라마&영화	정보	스포츠	교육	어린이	기타
1	0	1	0	0	0	0	0

송 컨텐츠를 추천하는 예이다.

가중치 적용 Markov 모델에 대한 장르의 선호도 전이는 표 4에서 보는 것처럼 식 (7)을 이용하여 계산할 수 있다. 표 4에서 아래에서 4번째 줄은 식 (8)을 이용하여 얻어진 장르의 가중치를 가리킨다. ρ = 1일 때, 표 5로부터 d-1 day(2003년 5월 6일) 시간구간 t에 시청한 장르는 뉴스와 드라마&영화 이므로, 이 두 장르에 대한 가중치 값(뉴스에 대한 가중치: 1.429, 드라마&영화에 대한 가중치: 1.611)만 선호도 전이 확률 값을 계산하는데 적용될 수 있다. 계산 결과 뉴스와 드라마&영화가 추천될 수 있다. 그 결과 가중치 적용 Markov 모델의 정확도는 100%이다.

이와 같은 예로부터 가중치 적용 Markov 모델을 이용한 컨텐츠 추천이 전통적인 Markov 모델에 비해 정확도가 높을 수 있음을 하나의 예로 살펴 보았다. 그 이유는 TV 시청자들은 전날 이미 보았던 선호하는 TV 프로그램을 일반적으로 시청하려 하기 때문이다. 그러므로 장르가 연속적으로 시청되면 될수록, 향후 선호도 전이를 추정하는데 더 많은 영향을 미칠 수 있다.

5. 성능 평가

본 논문에서 제안한 가중치 적용 Markov 모델의 성능 평가를 위해 실험 데이타로는 한국의 대표적인 시장조사 기관 중 하나인 AC Nielsen Korea에 위해 2002년 12월 1일부터 2003년 5월 31일까지 2,518명의 TV 시청자로부터 수집된 3,199,990건의 TV 시청 데이타를 이용하였다. TV 시청 데이타는 각 고객의 가정에 설치된 Set-Top Box를 이용하여 로그 인, 로그 아웃 시간, 방송 시간과 요일, 시청 프로그램의 장르 등을 수집하였다. 실험 데이타가 가지는 각 프로그램은 8개의 장르, 즉 뉴스, 오락, 드라마&영화, 정보, 스포츠, 교육, 어린이, 기타로 구분한다. 성능 평가를 위한 실험 환경은 인텔 Pentium IV, CPU 2.4 GHz, 메모리 512 Mbytes 환경에서 Java를 이용하여 구현하였다. 운영체제로는 윈도우 XP, 데이타베이스는 Microsoft Access 2000을 이용하였다.

그림 5는 본 논문에서 제안하는 가중치 적용 Markov 모델을 이용하여 추천된 프로그램 정보를 보여주는 방송 컨텐츠 추천 프로토타입 시스템의 실행 화면이다. 그림 5에서 보이는 것처럼, 예를 들어 News 프로그램을 시청하는 동안 새로운 프로그램을 시청하기 위해 시청자가 개인화된 전자 프로그램 가이드(PEPG) 버튼을 선택하면TV 화면 오른쪽 위쪽에 팝업 창으로 나타낼 수 있다. 이렇게 함으로써 시청자는 기존의 전자 프로그램 가이드(EPG)를 이용하여 자신이 선호하는 TV 프로그램을 찾는데 걸리는 시간을 절약할 수 있다.

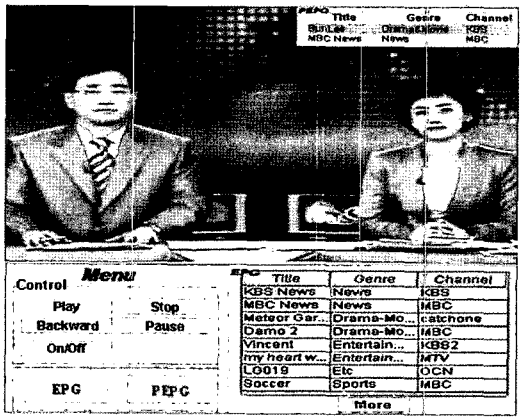

그림 5 방송 시청 중 PEPG 버튼 선택 화면

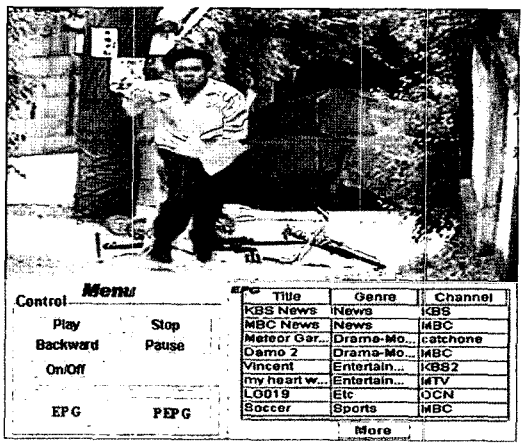

그림 6 추천된 프로그램 화면

그림 6은 가중치 적용 Markov모델에 의해 추천된 장르가 Drama인 방송 컨텐츠를 보여주고 있는 실행 화면 예이다.

5.1 성능 평가 방법

실험 데이타로 이용된 전체 6개월 데이타 중 처음 5개월 데이타는 훈련데이타로 이용하고, 나머지 1개월 데이타는 테스트 데이타로 이용하였다. 실험 데이타는 오

후 7시부터 11시까지 시청한 데이타를 1시간 간격으로 이용하였다. 이 시간 외의 데이타는 시청 빈도수가 매우 적기 때문에 제외하였다. 시험 데이타에 대한 시청자의 선호 장르를 예측하거나 추천하기 위한 최적의 파라미터를 찾기 위한 훈련 데이타는 길이의 일관성을 위해 시청자의 TV 시청 데이타 중 가장 오래된 날의 데이타는 제거하고 대신에 최근에 실험된 날의 데이타가 훈련 데이타에 포함된다.

성능 평가 대상은 TOP-N 모델[35,40], Markov 모델, 그리고 가중치 적용 Markov 모델에 대한 각각의 추천 정확도를 비교하였다. TOP-N 모델은 훈련 데이타에서 추천하고자 하는 특정 요일, 특정 시간대에 시청한 정보를 이용하여 8개의 장르 중 시청 빈도 수가 높은 순으로 추천하는 방법이다. Markov 모델은 시간에 따른 연속된 행위에서 바로 이전 행위를 기반으로 향후 행위를 예측할 수 있는 모델로 장르간에 시간 변화에 따른 선호도 전이 행렬을 이용한다. 즉 연속적으로 시청하는 TV 시청 프로그램을 기반으로 시간 변화에 따라 한 장르에서 다른 장르로 이동하는 확률을 고려한다. 가중치 적용 Markov 모델은 Markov 모델에 요일 변화에 따라 동일 TV 프로그램 장르를 연속해서 시청하는 경우 가중치를 부여하는 것으로 본 논문에서 제안하고 있는 모델이다. 시간 변화에 따른 통계적 선호 전이 행렬과 요일에 따라 연속적으로 시청하는 동일 장르에 대한 선호도 전이 행렬을 이용한다.

성능 평가 절차는 다음과 같다. 첫째, 가중치 적용 Markov 모델의 최적 파라미터 값을 추정한다. 고객이 방송 컨텐츠를 시청하고 있는 다음 방송 컨텐츠를 추천하기 위해서는 먼저, 식 (8)과 식 (11)에서와 같이 최적의 파라미터 값을 먼저 추정하여야 한다. 따라서, scaling factor ρ와 장르 j에 대해 2일 연속 시청에 대한 가중치 $w^{j}_{d-1,d}$, 그리고 Markov 모델의 다양한 차수 r에 대해 최적의 파라미터(ρ, r, k) 값을 찾는다. 그리고, 기존 Markov 모델과 파라미터 값의 변화에 따른 가중치 적용 Markov 모델의 성능의 차이를 비교 분석한다. 둘째, 다음 절에서 언급할 각각의 성능 평가 모델에 정확도를 비교한다. 정확도는 다음과 같이 계산할 수 있다.

정확도 = (실제 선호컨텐츠 수 ∩ 추천컨텐츠 수)*100 / 실제 선호컨텐츠 수

셋째, 훈련 데이타의 수집 기간을 변화 시켰을 때의 평균 정확도를 측정함으로써 훈련 데이타 량에 따라 분석 모델의 성능에 어떤 영향을 미치는지 분석한다.

5.2 성능 평가 결과 및 분석

가중치 적용 Markov 모델의 최적 파라미터 값을 추정하기 위해 파라미터 값(ρ, r, k)의 변화에 따른 가중치 적용 Markov모델의 정확도를 비교하였다. 여기서 ρ는 경험적 실험에 의해서 최적의 값을 선정할 수 있는 scaling factor이다. r은 Markov 모델의 차수, k는 $d-k$ 와 d day 사이에 2번 연속해서 사용하는 경우 가중치를 부여하기 위한 파라미터이다.

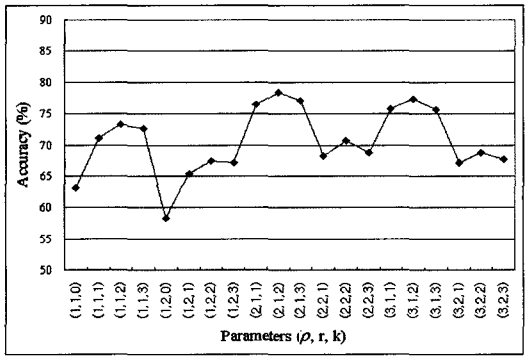

그림 7 가중치 적용 Markov 모델의 파라미터 값의 변화에 따른 정확도

그림 7은 가중치 적용 Markov 모델의 파라미터 값의 변화에 따른 정확도이다. 파라미터 값이 (ρ = 1, r = 1, k = 0)인 경우가 1차 Markov 모델에 해당되는 것으로 평균 정확도는 63% 이며, 2차 Markov 모델은 (ρ = 1, r = 2, k = 0) 인 경우로 평균 정확도는 58% 정도로 1차 Markov 모델보다 정확도가 떨어짐을 보이고 있다. 이는 Markov 모델의 차수가 증가함으로써 패턴에 대한 조합이 증가함으로써 훈련 데이타가 충분하지 않을 때 발생할 수 있는 low coverage를 유발하기 때문이다. 그림 5에서 ρ값과 k 값이 1 이상일 때가 가중치 적용 Markov 모델에 해당되는 것으로, 가중치 적용 Markov 모델을 이용함으로써 정확도가 약 15%~20% 정도 향상 되었음을 알 수 있다. 파라미터 값이 (ρ = 2, r = 1, k = 2)일 때 78%로 가장 높은 정확도를 얻었다. 따라서 가중치 적용 Markov 모델의 최적 파라미터 값으로 ρ = 2, r = 1, k = 2를 선정하였다.

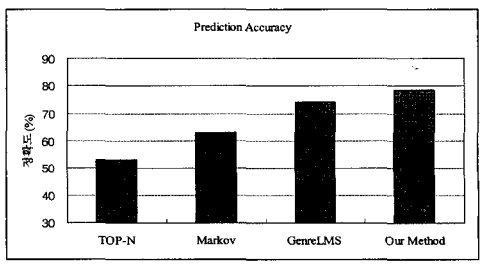

그림 8 평가 모델 간 성능 비교

본 논문에서 제안하는 모델의 성능을 평가하기 위하여, 제안하는 모델의 성능과 TOP-N 모델[35,40], Markov 모델[25-27], GenreLMS[36]의 성능을 비교하였다. 그림 8은 성능 평가 모델간 성능의 정도를 비교하기 위한 그래프이다. TOP-N 모델은 통계적 추정에 의해 시청 빈도수가 높은 순으로 추천하는 것으로 정확도가 약 53% 정도이다. Markov 모델의 경우 정확도는 약 63%이며, GenreLMS의 경우 약 74%이며, 제안 모델의 경우 약 78% 정도로 제안 모델의 정확도가 상대적으로 높음을 확인할 수 있다.

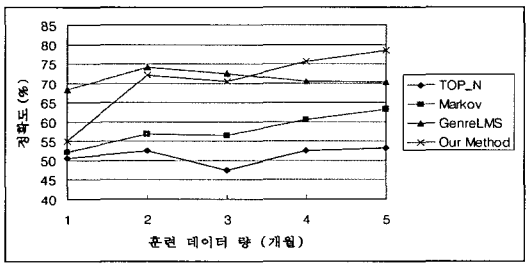

그림 9 훈련 데이타 기간 변화에 따른 평가 모델 간 성능 비교

그림 9는 훈련 데이타의 크기에 따른 평균 정확도를 측정한 그래프이다. 이 그래프를 보면, TOP-N 방법의 정확도가 다른 방법에 비해 데이타 크기에 관계없이 전체적으로 낮다. 그 이유는 TOP-N 방법은 시간에 따라 변하는 고객의 선호도 변화를 반영하지 못하기 때문이다. GenreLMS 방법은 데이타 크기가 작을 때(1개월)제안하는 방법보다 성능이 높다. 그 이유는 제안 모델에서 연속적으로 시청하는 프로그램에 제공하는 가중치를 추청하기 위한 정보가 데이타 량이 적을 경우 연속적으로 시청한 데이타가 희박하기 때문이다. 연속되는 데이타가 희박한 경우 계산된 가중치의 신뢰도가 떨어지는 원인이 된다. 제안 모델은 데이타 크기가 4개월 이상인 경우 GenreLMS보다 성능이 우수하다. 결과적으로 제안하는 가중치 적용 Markov 모델은 4개월 이상의 훈련 데이타를 이용하여 추천하는 것이 효과적임을 알 수 있다.

6. 결 론

본 논문에서는 시간에 따라 다양한 컨텐츠를 제공하는 방송 환경에서 시청자개인이 선호하는 방송 컨텐츠를 효율적으로 추천할 수 있는 가중치 적용 Markov 모델을 제안하고 성능 평가를 실시하였다. 전통적인 Markov 모델은 시간 변화에 따른 고객의 과거 컨텐츠 선호도 전이 행렬만을 이용한다. 방송 컨텐츠를 이용하는 고객들의 대부분은 시간 변화에 따라 이용되는 컨텐츠가 다양하며, 요일에 따라 연속해서 방송되는 컨텐츠는 고객이 관심을 가지는 컨텐츠라면 연속해서 시청할 가능성이 매우 높다. 따라서 본 논문에서는 요일에 따라 연속해서 방송되는 컨텐츠에 대한 고객의 관심도를 가중치로 계산하기 위한 알고리즘을 제안하였으며, 기존의 전통적인 Markov 모델에 가중치를 적용하였다. 제안된 알고리즘의 성능 평가를 위해 실제 고객이 이용했던 과거 시청 정보를 이용하여 본 논문에서 제안하는 방법과 TOP-N, 전통적인 Markov 모델, 그리고 GenreLMS의 정확도를 비교한 결과 전반적으로 우수한 성능을 보인다.

본 논문에서 제안한 가중치 적용 Markov 모델은 시간 변화에 따른 고객의 선호도 전이를 반영하는 방송 컨텐츠 환경에서 고객의 선호 컨텐츠를 추천하는 제한적인 방법을 사용하고 있으며, 고객 자신의 과거 데이타만을 이용하고 있다. 향후 과제에서는 다양한 환경에서 일반적인 멀티미디어 컨텐츠를 추천하기 위한 연구가 필요하며, 고객 자신의 과거 데이타가 부족할 경우 유사 고객 집단으로부터 정보를 활용함으로써 정확도를 높일 수 있는 연구를 진행할 예정이다.

참 고 문 헌

[1] Cotter, P., Smyth, B., "Personalization techniques for the digital TV world," Proc. Prestigious Applications of Intelligent Systems, 2000.

[2] S. Kang, J. Lim, and M. Kim, "Statistical Inference Method of User Preference on Broadcasting Content," LNCS, Vol. 3514, May 2005.

[3] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, "Probabilistic Memory-Based Collaborative Filtering," IEEE Transaction on Knowledge and Data Engineering, Vol. 16, No. 1, January 2004.

[4] Quinlan, J.R., "Induction of decision trees," Machine Learning, Vol. 1, No. 1, 1986.

[5] Michael J. A. Berry, Gordon Linoff, Data Mining Techniques For Marketing, Sales, and Customer Support, John Wiley&Sons, Inc., 1997.

[6] Kim, M., Ryu, G., Bae, B., "Intelligent program guide for digital broadcasting," Proc. International Workshop on Advanced Image Technology, 2002.

[7] Aggrawall, R., Imielinski, T., Swami, A., "Mining association rules between sets of items in large databases," Proc. ACM SIGMOD Int'l Conference on Management of Data, 1994.

[8] Aggrawall, R., Srikang, R., "Fast algorithms for mining association rules," Proc. 20th Int'l Conference on Very Large Databases, 1994.

[9] Ashrafi, M.Z., Tnizr, D., Smith, K., "An optimized distributed association rule mining algorithm," IEEE Distributed Systems Online, Vol. 5, No. 3, 2004.

[10] M. Balabanovic and Y. Shoham, "Fab: Content-Based Collaborative Recommendation," Communication ACM, Vol. 40, No. 3, 1997.
[11] Maltz, D.A., "Distributing information for collaborative filtering on Usenet net news," SM Thesis, Massachusetts Institute of Technology, 1994.
[12] Schafer, J.B., Konstan J., Riedl, J., "Recommender systems in e-commerce," ACM Conference on Electronic Commerce, 1999.
[13] Linden, G., Smith, B., York, J., "Amazon.com recommendations: item-to-item collaborative filtering," IEEE Internet Computing, 2003.
[14] Herlocker, J.L., Konstan, J.A., Terveen, L.G., Lidle J.T., "Evaluating collaborative filtering recommender systems," ACM Transactions on Information Systems, Vol. 22, No. 1, 2004.
[15] Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J., "Enhancing digital libraries with TechLens+," ACM/IEEE-CS Joint Conference on Digital Libraries, 2004.
[16] Cotter, P., Smyth, B., "Personalization techniques for the digital TV world," Proc. Prestigious Applications of Intelligent Systems, 2000.
[17] Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Chiarotto, A., Difino, A., Negro, B., "Personalized recommendation of TV programs," Lecture Notes in Artificial Intelligence, Vol. 2829, 2003.
[18] Ardissono, L., Gena, C., Torasso, P., "User Modeling and Recommendation Techniques for Personalized Electronic Program Guides, Personalized Digital Television," Targeting Programs to Individual Viewers Series: Human-Computer Interaction Series, Vol. 6, Kluwer Academic Publishers, 2004, http://www.wkap.nl/prod/b/1-4020-2163-1.
[19] Brown, S.M., Santos Jr., E. Banks, S.M., "A dynamic Bayesian intelligent interface agent," Proc. The Sixth International Interfaces Conference, 1997.
[20] Jensen, F.V., Bayesian Networks and Decision Graphs. Springer, 2001.
[21] Lin, F., Liu, W. Chen, Z., Whang, H., Tang, L., "User modeling for efficient use of multimedia files," Lecture Notes in Computer Science, Vol. 2175, 2001.
[22] Morris, Q., "Recognition networks for approximate inference in BN2O networks," Proc. The seventeenth Conference on Uncertainty in Artificial Intelligence, 2001.
[23] Ng, A.Y., Jordan, M.I., "Approximate inference algorithms for two-layer Bayesian networks," In Advances in Neural Information Processing Systems, 2000.
[24] Frey, B.J., Patrascu, R., Jaakkola, T.S., Moran, J., "Sequentially fitting inclusive trees for inference in noisy-OR networks," In Advances in Neural Information Processing Systems, Vol. 13, 2000.
[25] Sarukkai, R. R., "Link prediction and path analysis using Markov chains," The International Journal of Computer and Telecommunications Networking, Vol. 33, 2000.
[26] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S., "Visualization of navigation patterns on a web site using model based clustering," Proc. International KDD Conference, 2000.
[27] He, S., Qin, Z., Chen, Y., "Web pre-fetching using adaptive weight hybrid-order markov model," Lecture Notes in Computer Science, Vol. 3306, 2004.
[28] P. Resnick and H.R. Varian, "Recommender Systems," Communications of the ACM, Vol. 40, No. 3, Mar. 1997.
[29] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer, 2000.
[30] P. Cotter and B. Smyth, "A Personalized Television Listing Service," Communications of the ACM, Vol. 43, No. 8, Aug. 2000.
[31] http://www.ptv.ie/
[32] W.P. Lee and J.H. Wang, "A User-Centered Remote Control System for Personalized Multimedia Channel Recommendation," IEEE Transactions on Consumer Electronics, Vol. 50, No. 4, Nov. 2004.
[33] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlock, L.R. Gordon., and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," Communications of The ACM, Vol. 40, No. 3, Mar. 1997.
[34] L. Ardissono, F. Portis, P. Torasso, F. Bellifemine, A. Chiarotto, and A. Difino, "Architecture of a System for the Generation of Personalized Electronic Program Guides," Workshop on Personalization in Future, 2001., http://www.di.unito.it/~liliana/
[35] H.K.Lee, J.Nam, B. Bae, M. Kim, K. Kang, J. Kim, "Personalized Contents Guide and Browsing Based on User Preferece," Proc. the AH2002 Workshop on Personalization in Future TV, pp. 130-137, 2002.
[36] Setten M.V., "Experiments with a Recommendation Technique that Learns Category Interests," Proc. IADIS WWW/Internet, pp. 722-725, 2002.
[37] S. Kang, J. Lim, and M. Kim, "Modeling the user preference of broadcasting content using Bayesian networks," Journal of Electronic Imaging, Vol. 14(2), Apr. 2005.
[38] Mukund Deshpande and George Karypis, "Selective Markov Models for Predicting Web-page," ACM Transactions on Internet Technology (TOIT) Vol. 4 , Issue 2, May 2004.
[39] Xing Dongshan and Shen Junyi, "a New Markov Model for Web Access Prediction," Computing in Science & Engineering, November-December 2002.
[40] Mukund D., George K., "Item-Based Top-N

Recommendation Algorithms," ACM Transactions on Information System, Vol. TBD, TBD 20 TBD, 2004.

박 성 준

1985년 동국대학교 통계학과 학사. 1987년 동국대학교 통계학과 석사. 2001년 충남대학교 컴퓨터과학과 석사. 2005년 충남대학교 컴퓨터과학과 박사. 1989년~1998년 ETRI 선임연구원. 2002년~현재 공주영상대학 모바일게임과 조교수. 관심분야는 개인화, 멀티미디어, 데이타마이닝, 무선인터넷, 유비쿼터스

홍 종 규

2004년 충남대학교 컴퓨터과학과 학사 2006년 충남대학교 컴퓨터공학과 석사 2006년~현재 (주)이노플러스 연구원. 관심분야는 지능형 추천 알고리즘, e-비지니스

강 상 길

1989년 성균관대학교 전기공학과 학사 1995년 Columbia University, 석사 2002년 Syracuse University 박사. 2006년~현재 인하대학교 컴퓨터공학부 조교수. 관심분야는 멀티미디어, 개인화, 유비쿼터스, 인공지능, 데이타마이닝

김 영 국

1985년 서울대학교 계산통계학과 학사 1987년 서울대학교 계산통계학과 석사 1995년 버지니아대 컴퓨터과학과 박사 1995년~1996년 핀란드 VTT, 노르웨이 SINTEF DELAB 방문연구원. 1996년~현재 충남대학교 전기정보통 공학부 부교수. 2002년 8월~2003년 7월 UC Davis 방문교수. 관심분야는 실시간 데이타베이스, 모바일정보시스템, 전자상거래 시스템