퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 2, pp. 164-171

선호도 전이 확률을 이용한 멀티미디어 컨텐츠 추천 시스템

A Multimedia Contents Recommendation System using Preference Transition Probability

박성준* · 강상길** · 김영국***

Sungjoon Park, Sanggil Kang and Young-Kuk Kim

* 공주영상대학 모바일게임과

** 수원대학교 컴퓨터학과

*** 충남대학교 전기정보통신공학부 컴퓨터전공

요 약

최근에 서비스되기 시작한 디지털 멀티미디어 방송은 다양한 종류의 수많은 컨텐츠를 제공하기 때문에 고객은 때로 자신이 선호하는 컨텐츠를 찾는데 많은 시간을 소비한다. 심지어는 선호 컨텐츠를 찾는 동안 이미 방송이 끝날 수도 있다. 이와 같은 문제를 해결하기 위해서는 고객이 필요로 하는 최소 정보만을 추천하기 위한 방법이 필요하다. 본 논문에서는 고객이 시청한 컨텐츠 선호도 전이 확률을 이용하여 고객이 선호하는 컨텐츠를 미리 예측하여 추천하기 위한 알고리즘과 시스템을 제안한다. 제안하는 시스템은 클라이언트 관리자 에이전트, 모니터링 에이전트, 러닝 에이전트, 그리고 추천 에이전트 모듈로 구성된다. 클라이언트 관리자 에이전트는 다른 모듈과 상호 작용을 하면서 조정자 역할을 한다. 모니터링 에이전트는 컨텐츠에 대한 고객의 선호도를 분석하기 위해 고객이 이용했던 usage history 데이터를 수집하기 위한 에이전트이다. 러닝 에이전트는 고객으로부터 수집된 usage history 데이터를 정제하여 시간 변화에 따른 상태 전이 행렬로 모델링하기 위한 에이전트이다. 추천 에이전트는 고객의 상태 전이 행렬로 구성된 모델링 데이터에 본 논문에서 제안하는 선호도 전이 확률 모델을 이용하여 고객이 바로 다음에 선호하게 될 컨텐츠를 추천하기 위한 에이전트이다. 추천 에이전트 모듈에서 컨텐츠에 대한 고객의 선호도 전이 확률을 이용하는 추천 알고리즘을 제안한다. 제안하는 추천 시스템은 무선 인터넷 표준 플랫폼인 WIPI(Wireless Internet Platform for Interoperability) 플랫폼에서 프로토타입 시스템을 설계, 구현하였으며, 실험결과 제안된 선호도 전이 확률 모델의 추천 정확도가 전형적인 방법에 비해 효과적임을 보인다.

Abstract

Recently Digital multimedia broadcasting (DMB) has been available as a commercial service. The users sometimes have difficulty in finding their preferred multimedia contents and need to spend a lot of searching time finding them. They are even very likely to miss their preferred contents while searching for them. In order to solve the problem, we need a method for recommendation users preferred only minimum information. We propose an algorithm and a system for recommending users'preferred contents using preference transition probability from user's usage history. The system includes four agents: a client manager agent, a monitoring agent, a learning agent, and a recommendation agent. The client manager agent interacts and coordinates with the other modules, the monitoring agent gathers usage data for analyzing the user's preference of the contents, the learning agent cleans the gathered usage data and modeling with state transition matrix over time, and the recommendation agent recommends the user's preferred contents by analyzing the cleaned usage data. In the recommendation agent, we developed the recommendation algorithm using a user's preference transition probability for the contents. The prototype of the proposed system is designed and implemented on the WIPI(Wireless Internet Platform for Interoperability). The experimental results show that the recommendation algorithm using a user's preference transition probability can provide better performances than a conventional method.

Key words : 추천시스템, 개인화, 모바일, 멀티미디어 컨텐츠

1. 서 론

최근 무선 인터넷 및 모바일 단말기의 발전은 사용자 수의 증가와 함께 게임, 뉴스, 음악, 디지털 방송 등과 같은 다양한 종류의 서비스를 언제, 어디서나 쉽게 제공할 수 있게 되었으며, 소비되는 정보의 양이 기하급수적으로 증가하고 있다. 이와 같은 다양한 많은 정보를 효율적으로 활용하기 위해서는 고객이 원하는 정보를 언제, 어디서나 쉽게 이용할 수 있도록 하는 것이 중요하다.

지금까지 아마존, CDNOW 등과 같은 인터넷 쇼핑몰 업체들은 유사 고객들의 반응을 기반으로 고객이 만족할 만한 상품이나 광고를 추천해줌으로써 상업적으로도 많은 성공을 거

접수일자 : 2006년 1월 25일
완료일자 : 2006년 4월 12일
감사의 글 : 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성 지원 사업의 연구결과로 수행되었음(IITA-2005-C1090-0502-0016).

두고 있으며, 웹 사이트 개인화를 통해 고객 개인에 맞는 상품이나 광고를 제공하기 위한 많은 연구[1-7]가 진행되고 있다. 그러나 이들 연구의 대부분은 여러 고객들의 과거 소비 행위 또는 고객이 직접 시스템에 입력하는 피드백 정보를 이용하여 목표 고객과 유사 특성을 갖는 고객들의 반응을 기반으로 고객이 만족할 만한 상품이나 광고를 추천해 주는 방식이다. 이와 같은 방식은 여러 고객으로부터 정보를 수집하는데 많은 시간, 비용, 노력이 요구되며, 수집된 정보의 량이 과다하고 많은 리소스를 필요로 한다.

이에 비해 모바일 환경은 적은 화면, 불편한 사용자 인터페이스, 낮은 성능, 네트워크의 불안정성과 높은 비용 등 부족한 리소스를 가지는 모바일 단말기를 통해 언제, 어디서나 컨텐츠를 이용하게 되며, 이용되는 컨텐츠의 유형도 시간 변화에 따라 다양하게 전이되는 특징을 가진다. 이와 같이 모바일 단말기가 가지는 제약 사항들의 한계를 극복하고, 시간 변화에 따른 컨텐츠 소비 성향이 변화는 환경에서 고객이 원하는 정보를 언제 어디서나 빠른 시간에 이용할 수 있는 방법이 절실히 요구되고 있다.

디지털 멀티미디어 방송(DMB: Digital Multimedia Broadcasting)은 휴대폰, PDA(Personal Digital Assistants), 휴대용 TV 등과 같은 모바일 장비들을 이용하여 텍스트, 오디오, 비디오 등과 같은 방송 컨텐츠를 언제, 어디서나 제공할 수 있다. 그러나 DMB 방송은 컨텐츠 수가 엄청나게 많기 때문에 TV 시청자들은 때로 자신이 선호하는 프로그램을 찾는데 많은 시간을 소비한다. 심지어는 선호 프로그램을 찾는 동안 이미 방송이 끝날 수도 있다. 이와 같은 환경에서 고객이 선호하는 컨텐츠를 미리 예측하여 추천함으로써 어느 정도 보다 편리한 생활을 고객에게 제공할 수 있다.

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. 개인화 방송은 고객이 선호하는 프로그램을 찾는데 걸리는 시간을 줄여줄 수 있으며, 선호 프로그램을 찾는 동안 이미 방송이 진행되어 원하는 방송을 놓쳐버리는 경우를 줄여줌으로써 보다 편리한 생활을 제공할 수 있다. 또한, 개인화 방송은 선호하는 컨텐츠만을 제공할 수 있기 때문에 모바일 장치의 대역폭을 효율적으로 사용하는데 기여할 수 있다.

모바일 환경은 적은 화면, 낮은 대역폭, 낮은 성능, 리소스의 한계, 불편한 사용자 인터페이스 등 모바일 장치가 가지는 여러 가지 제약 사항들이 있다. 이를 극복하기 위해서는 서버와의 통신 횟수를 줄이고, 이용되는 리소스를 최소화할 수 있는 추천 시스템을 제공할 필요가 있다. 따라서 본 논문에서는 모바일 환경이 가지는 제약사항을 극복하면서, 방송 컨텐츠 특성 상 시간 변화에 따라 컨텐츠 소비 성향이 전이되는 고객의 선호도 예측을 위해 제안된 선호도 전이 확률 모델을 이용하여 클라이언트 쪽에서 추천 알고리즘이 수행되는 방송 컨텐츠 추천 시스템을 제안한다. 추천 시스템은 모니터링 에이전트(Monitoring Agent), 러닝 에이전트(Learning Agent), 그리고 추천 에이전트(Recommendation Agent)로 구성 된다. 모니터링 에이전트는 시청자 컨텐츠 선호도 전이를 분석하기 위해 최근에 시청한 컨텐츠 정보를 수집한다. 러닝 에이전트는 고객의 선호도를 분석하기 위한 기본 자료를 만들기 위해 수집된 자료를 정제하고 선호도 전이 행렬로 모델링한다. 추천 에이전트는 러닝 에이전트에 의해 정제된 고객의 최근 시청 정보를 본 논문에서 제안하는 추천 알고리즘을 이용하여 추론 값이 높은 순으로 컨텐츠를 추천한다. 또한, 본 논문에서는 무선 인터넷 표준 플랫폼인 WIPI(Wireless Internet Platform for Interoperability)[8] 플랫폼 상에서 제안하는 개인화 컨텐츠 추천 프로토타입 시스템을 설계하고 구현하였다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련 연구에 관하여 서술하고, 제3장에서는 고객이 선호하는 컨텐츠를 추천하기 위한 멀티미디어 컨텐츠 추천 시스템에 대해 기술한다. 제4장에서는 본 논문에서 제안하는 멀티미디어 컨텐츠 추천 시스템의 구현 및 실험 결과를 기술하고, 마지막으로 제5장에서 결론을 맺는다.

2. 관련 연구

멀티미디어 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[9]이고, 다른 하나는 추론 방식[10]이다. 협업 필터링은 고객과 비슷한 소비 성향을 가지는 고객 그룹을 기반으로 고객이 선호로 하는 컨텐츠를 결정한다. 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 이루어진다. 서버 쪽 개인화 시스템의 예로는 Cotter et al.[11]이 TV 프로그램에 대한 고객의 명시적 선호 점수를 기반으로 하는 웹 기반 개인화 전자 프로그램 가이드인 PTV[12]를 소개한다. Lee et al.[13]는 서버 쪽에서 자동화된 컨텐츠 추천을 위한 고객 중심의 리모콘 시스템을 데모한다. Resnick et al.[14]과 Konstan et al.[15]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[16-20]. 클라이언트 쪽에서는 고객의 사용 정보만 추론하는데 필요하기 때문에, 클라이언트 쪽에서 고객이 선호하는 컨텐츠를 추론하기 위한 데이터의 양은 서버 쪽의 데이터 양에 비해 상대적으로 아주 작다. 클라이언트 쪽에서 개인화의 가장 대표적인 예는 디지털 방송 시스템에서 개인화된 전자 프로그램 가이드(PEPG: Personalized Electronic Program Guide)이다. Kang et al.[16-19]는 TV 장르에 대한 현재 선호도와 과거 선호도간의 상호 정보를 이용하여 TV 장르를 추천하기 위한 알고리즘을 소개하였다.

3. 목표 시스템 구조

3.1 전체 시스템 구조

본 연구에서 제안하는 선호도 전이 확률을 이용한 멀티미디어 컨텐츠 추천 시스템 구조는 그림 1과 같이 구성된다. 제안하는 시스템은 클라이언트 쪽에 클라이언트 관리자 에이

전트(Client Manager Agent), 모니터링 에이전트(Monitoring Agent), 러닝 에이전트(Learning Agent), 그리고 추천 에이전트(Recommendation Agent)로 구성된다.

클라이언트 관리자 에이전트는 다른 에이전트들과의 상호작용을 하면서 조정자 역할을 한다. 모니터링 에이전트는 컨텐츠에 대한 고객의 선호도를 분석하기 위해 고객이 이용했던 usage history 데이터를 수집하기 위한 에이전트이다. 수집된 데이터는 여러 가지 이유로 쓰레기 데이터를 가지고 있을 수 있기 때문에 수집된 데이터를 정제할 필요가 있다. 러닝 에이전트는 고객으로부터 수집된 usage history 데이터를 정제하여 시간 변화에 따른 상태 전이 행렬로 모델링하기 위한 에이전트이다. 추천 에이전트는 고객의 상태 전이 행렬로 구성된 모델링 데이터에 본 논문에서 제안하는 선호도 전이 확률 모델을 이용하여 고객이 바로 다음에 선호하게 될 컨텐츠를 추천하기 위한 에이전트이다.

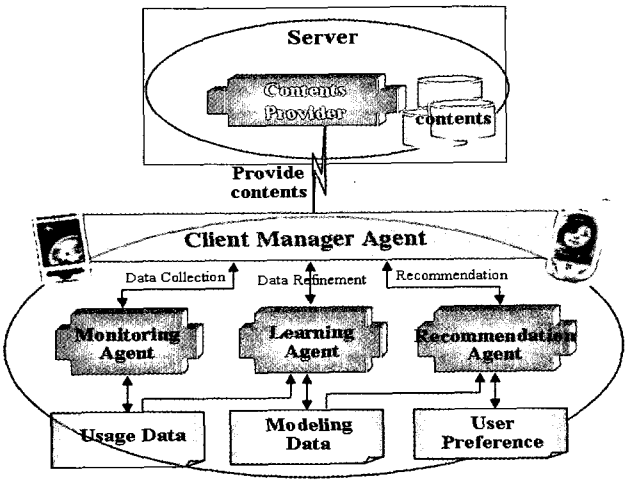

그림 1. 멀티미디어 컨텐츠 추천 시스템 구조.
Fig. 1. Architecture of Multimedia Contents Recommendation System.

컨텐츠 제공자는 그림 2와 같이 제목, 장르, 채널, 주인공 등과 같은 컨텐츠의 주요 정보를 설명하는 메타 데이터와 함께 컨텐츠를 클라이언트 쪽에 제공한다. 메타 데이터는 고객이 선호하는 컨텐츠를 추천하기 위해 고객의 소비 행위를 분석하는데 필요한 주요 데이터이다.

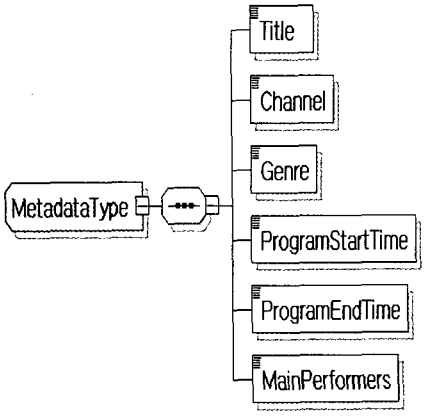

그림 2. 메타 데이터.
Fig. 2. Meta Data.

만약 고객이 특정 컨텐츠를 선택하면, 선택된 컨텐츠가 모바일 디바이스의 화면에 나타나게 되며, 컨텐츠와 관련된 메타 데이터가 클라이언트 관리자 에이전트에 의해 모니터링 에이전트에 보내진다. 모니터링 에이전트는 방영 시간 등과 같이 소비된 컨텐츠에 대한 고객의 소비 행위를 모니터링하고 이를 기록한다. 러닝 에이전트는 수집된 원시 데이터를 정제하고 추천 에이전트에서 원하는 형태로 결합하여 시간에 따른 선호도 상태 전이 행렬로 데이터를 모델링 한다. 그러면 추천 에이전트는 모델링 데이터를 이용하여, 본 논문에서 제안하고 있는 추천 알고리즘으로부터 고객이 선호하는 컨텐츠를 추론해서 XML 파일에 저장한다. 고객이 모바일 장치에 있는 개인화 컨텐츠 추천 버튼을 선택하면, 클라이언트 관리자 에이전트의 요청에 의해 추천된 컨텐츠 목록이 화면에 나타난다.

그림 3은 본 논문에서 제안하는 개인화 추천 시스템을 구성하는 클라이언트 관리자 에이전트와 다른 에이전트간의 연관성을 보여주는 에이전트 클래스들 간의 관계도이다. 고객은 휴대폰, PDA 등과 같은 다양한 모바일 장치들을 소유할 수 있으며, 클라이언트 쪽의 클라이언트 관리자 에이전트는 서버 쪽의 컨텐츠 제공자로부터 1개 이상의 컨텐츠를 제공받을 수 있다. 클라이언트 관리자 에이전트는 상황에 따라 모니터링 에이전트, 러닝 에이전트, 그리고 추천 에이전트를 호출할 수 있다.

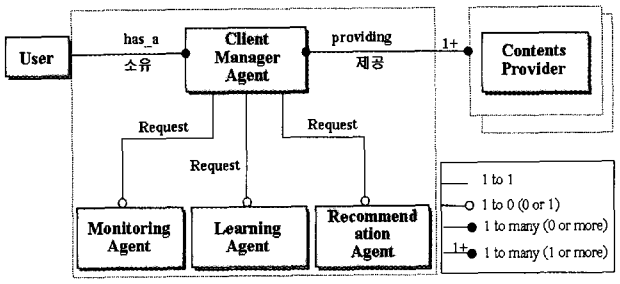

그림 3. 에이전트 클래스간의 관계도.
Fig. 3. Relationship between Agent class.

그림 4는 본 논문에서 제안하는 추천 시스템에서 요구되는 엔터티와 객체들 사이의 연관성 보여주기 위한 EER(Enhanced Entity-Relationship) 모델이다. 고객과 컨텐츠 간에는 usage history 데이터와 voting 관계를 통하여 고객은 자신이 원하는 컨텐츠를 시청하고, 시청 결과에 대하여 평가라는 피드백 정보를 남김으로써 보다 정확하게 자신이 원하는 추천 정보를 받을 수 있다.

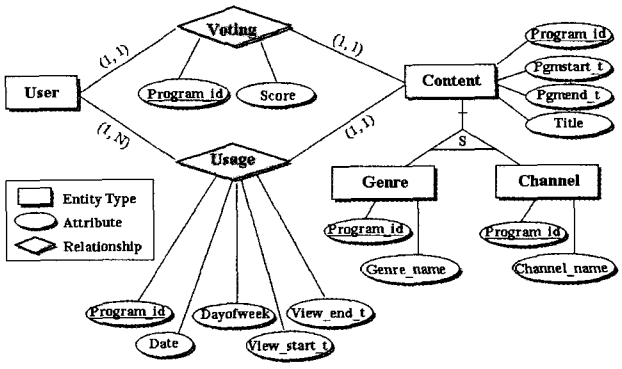

그림 4. EER 모델.
Fig. 4. EER Model.

그림 4에서 S+는 상위 클래스인 content는 하위클래스인 genre와 channel의 공통적인 속성(attribute)을 가지며, 하위 클래스인 genre와 channel은 각각 상위 클래스의 주 키(primary key)와 하위 클래스간의 공유되지 않는 속성을 갖는다. 주 키는 상위 클래스와 하위클래스 사이의 연결을 위해 필요하다. 상위 클래스인 content에 있는 모든 속성은 하위클래스인 genre와 channel에 상속된다.

3.2 클라이언트 관리자 에이전트 (Client Manager Agent)

클라이언트 관리자 에이전트는 클라이언트 쪽에 있는 다른 여러 에이전트들과 상호작용을 하면서 조정자 역할을 한다. 고객이 자신의 모바일 장치에 있는 DMB 모드에 로그인할 때, 클라이언트 관리자 에이전트는 추천 에이전트에게 개인화된 컨텐츠의 추천을 요청한다. 그러면 추천 에이전트는 선호 컨텐츠 목록이 보여지는 팝업 윈도우를 통해 고객이 선호하는 컨텐츠를 추천한다.

만약 고객이 추천한 컨텐츠 중 하나를 선택하면, 클라이언트 관리자 에이전트가 직접 컨텐츠를 연결하여 모바일 장치의 화면에 선택된 컨텐츠를 방영한다. 클라이언트 관리자 에이전트는 또한 컨텐츠가 끝나거나 고객의 마음이 변했을 때 소비된 컨텐츠의 메타데이터와 고객의 소비 행위를 기록하도록 모니터링 에이전트에게 요청한다. 고객이 DMB 모드를 종료하면 클라이언트 관리자 에이전트는 이용 데이터를 기록하도록 모니터링 에이전트에게 요청하고, 추천 에이전트가 데이터를 분석하기 좋은 형식으로 정제하여 모델링 데이터를 만들도록 러닝 에이전트에게 요청한다.

만약 고객이 소비한 컨텐츠에 대한 평가를 임의의 시점에 직접 모바일 장치에서 입력할 수 있으면, 클라이언트 관리자 에이전트는 선호 컨텐츠를 추천하는데 주요 정보가 될 수 있는 선호 점수를 기록 하도록 모니터링 에이전트에게 요청한다. 그러나 소비한 컨텐츠에 대한 선호도 점수를 입력하도록 하는 것은 때로는 고객이 싫어할 수 있다. 따라서 우리의 추천 알고리즘은 고객이 직접 입력한 컨텐츠에 대한 선호도 점수는 제외하고 고객의 컨텐츠 소비행위에 의해 자동으로 수집된 데이터만을 이용하여 고객의 개인화된 추천 컨텐츠를 선정하는데 초점을 두었다.

3.3 모니터링 에이전트 (Monitoring Agent)

모니터링 에이전트는 고객의 이용 데이터(usage data)를 수집하고, 수집하는 과정에서 발생할 수 있는 쓰레기 데이터를 정제하는 역할을 한다. 모니터링 에이전트는 메타데이터, 묵시적 데이터(implicit data), 그리고 명시적 데이터(explicit data)로 구성된 고객의 usage history 데이터를 수집하는 역할을 한다. 메타데이터는 서버 쪽 방송 시스템의 컨텐츠 제공 에이전트에 의해 제공되는 정보로 제목, 채널, 장르, 프로그램 시작 시간, 프로그램 종료 시간, 주인공 등과 같은 TV 프로그램을 이해할 수 있는 정보로 구성된다. 묵시적 데이터는 프로그램 시청 시간, 프로그램 시청 종료 시간 등과 같은 고객의 소비 행위에 의해서 수집될 수 있다. 묵시적 데이터는 소비된 TV 프로그램에 대한 시청 일자, 시청 요일, 시청 시작 시간, 시청 종료 시간을 포함한다. 명시적 데이터는 고객에 의해 직접 입력된 선호 점수(voting score)라고 부르는 소비된 컨텐츠에 대한 평가를 의미한다. 앞 절에서 언급하였듯이 명시적 데이터는 프로그램이 소비될 때마다 각 프로그램에 대한 자신의 선호도를 입력시키기 위해서는 고객의 노고와 시간이 소요된다. 따라서 본 논문에서는 명시적인 데이터는 배제하고 메타데이터와 묵시적인 데이터만을 이용하여 본 논문에서 제안하는 컨텐츠 추천 알고리즘을 구현한다.

3.4 러닝 에이전트 (Learning Agent)

모델링 데이터를 구성하기 위한 러닝 에이전트 구조는 그림 5와 같다. 모니터링 에이전트에 의해 수집된 이용 데이터는 수집되는 동안 쓰레기 데이터가 포함될 수 있다. 만약 컨텐츠를 소비한 시간이 너무 짧다면, 고객이 선호도를 가지고 컨텐츠를 소비했는지 의심스럽다. 예를 들어, 만약 방영 시간이 1시간인 TV 프로그램을 단지 1분 동안만 시청하였다면, 이 프로그램은 고객이 좋아서 시청한 것이라 생각할 수 없다. 러닝 에이전트는 전체 방영 시간에 비해 시청 시간의 비율이 너무 작은 컨텐츠는 제거한 후, 분석을 위한 모델링 데이터를 생성한다. 선호 컨텐츠를 추천하기 위해 컨텐츠를 과거 이용 데이터로부터 제외할 것이지 포함시킬 것이지 결정하기 위한 전환점(threshold)은 추천의 정확도를 어느 정도 향상시킬 수 있는 범위에서 결정될 수 있다. 최적의 전환점은 실제 TV 프로그램 이용 데이터를 이용하여 실행한 실제 실험에 의해 얻어질 수 있다. 또한 컨텐츠를 이용하는 중간에 예기치 못한 시스템 에러로 컨텐츠 시청 시작 시간만 로그 파일에 기록될 수 있다. 이와 같이 시청 시작 시간만 기록되어 있고 시청 종료 시간이 없는 로그 정보는 고객이 시청한 기간을 계산할 수 없으므로, 러닝 에이전트는 컨텐츠의 시청 정보를 제거한다. 이와 같이 정제된 usage history 데이터를 이용하여 현재 소비하고 있는 멀티미디어 컨텐츠 종류를 입력으로 받아 usage history 데이터로부터 시간 슬롯 t에서 $t+1$로의 통계적 전이 테이블(Transition Table)을 생성하며, 이 테이블을 저장한 것이 모델링 데이터이다.

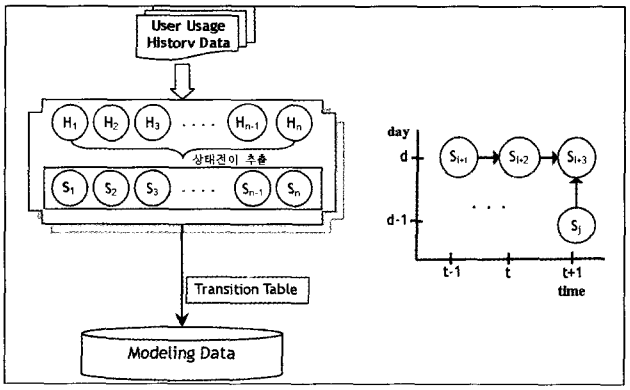

그림 5. 러닝 에이전트 구조.
Fig. 5. Architecture of Learning Agent.

3.5 추천 에이전트 (Recommendation Agent)

추천 에이전트 구조는 그림 9와 같다. 추천 에이전트는 러닝 에이전트에 의해 모델링된 고객의 최근 컨텐츠 선호도 전이 정보를 본 논문에서 제안하는 추천 알고리즘을 이용하여 추론 값이 높은 순으로 컨텐츠를 추천한다. 추천 에이전트는 아주 가까운 미래, 즉 바로 다음 타임에 고객이 선호하는 컨텐츠를 추천하기 위해 러닝 에이전트에서 사전에 정의된 기간 동안 수집되어 정제된 모델링 데이터를 이용한다. 과거 이용 데이터에서 현재 타임 슬롯에서의 선호도로부터 바로 다음 타임 슬롯에서의 선호도로의 컨텐츠에 대한 고객의 선호도 전이 정보는 가까운 미래에 선호로 하는 컨텐츠를 결정하기 위한 주요 요인이 될 수 있다.

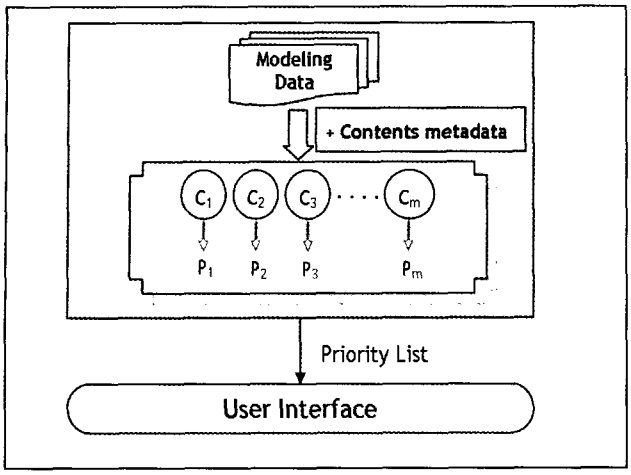

그림 6. 추천 에이전트 구조.
Fig. 6. Architecture of Recommendation Agent.

과거 일정 기간 동안의 선호도 전이 누적 테이블로부터 우리는 식(1)에서 정의된 것처럼 usage history 데이터를 이용하여 time slot t 에서 $content_i$가 소비되었다는 조건에서 모든 가능한 전이로 time slot $t+1$에서 소비된 $content_j$ 의 빈도수에 대한 비율을 계산함으로써 time slot t 에서 time slot $t+1$로 컨텐츠의 통계적 선호도 전이를 계산할 수 있다.

$$p(content_{i,t}, content_{j,t+1}) = \frac{T(i,j)}{\sum_{j=1}^{n} T(i,j)} \quad (1)$$

여기서, $content_{i,t}$ 는 time slot t 에서 소비된 contenti 이며, contentj,t+1 는 time slot $t+1$에서 소비된 $content_j$ 이다. 또한 $T(i,j)$는 time slot t 에서 소비된 $content_i$로부터 time slot $t+1$에서 소비된 $content_j$로의 전이 빈도수 이다. n은 컨텐츠의 수이다.

일반적으로 고객은 최근에 소비한 컨텐츠를 다시 소비하는 경향을 가지고 있다. 이와 같은 상황을 고려하기 위해서 본 논문에서는 식(2)와 같이 통계적 전이를 계산하는데 2일간 계속해서 연이어서 소비된 컨텐츠의 빈도수에 가중치를 부여한다.

$$p_\alpha(content_{i,t}, content_{j,t+1}) = \frac{w_{d-1,d}(j,j)T(i,j)}{\sum_{j=1}^{n} w_{d-1,d}(j,j)T(i,j)} \quad (2)$$

여기서,

$$w_{d-1,d}(j,j) = \left(1 + \frac{n_w(t+1)}{N_w(t+1)}\right) \quad (3)$$

여기서 $Nw(t+1)$는 usage history에 속하는 전체 주(weeks)를 의미한다. $nw(t+1)$는 $content_j$가 usage history에서 연속해서 2일간 time slot $t+1$에서 연속해서 소비된 횟수이다.

멀티미디어 컨텐츠 추천 알고리즘이 어떻게 수행되는지 예를 들어 알아보고자 한다. usage history로 이용될 데이터는 AC Nielson Korea로부터 제공된 6개월간의 실험 데이터 중 1개월간의 데이터 2002년12월7일부터 2003년1월6일까지의 데이터를 훈련 데이터로 이용하였다. 컨텐츠 추천 과정 및 추천의 정확도는 다음과 같은 절차에 의해 구할 수 있다.

1) 고객으로부터 수집된 usage history 데이터부터 추천하고자 하는 요일(day d)과 같은 요일의 시간 t에서 시청한 컨텐츠와 $t+1$에서 시청한 컨텐츠에 대한 통계적 선호도 전이 행렬을 모델링 한다.

2) usage history 데이터부터 day $d-1$, $t+1$ 시간에 시청한 컨텐츠와 day d, $t+1$시간에 시청한 컨텐츠와 동일한 컨텐츠에 가중치 계산 알고리즘을 이용하여 가중치를 계산한다.

3) 1), 2)로부터 얻은 통계적 선호도 전이 행렬과 가중치를 이용하여 각각의 컨텐츠에 대한 확률 값을 계산한다.

4) 3)에서 계산된 확률 값으로부터 최상위 값을 가지는 2개의 선호 컨텐츠를 추천한다. 여기서 2개를 추천하는 이유는 수집 데이터로부터 1시간에 시청하는 평균 컨텐츠 수가 약 2개이기 때문이다.

5) 테스트 데이터에서 d day, 시간 구간 $t+1$에 시청한 실제 장르와 추천된 장르를 비교하여 일치하는 장르의 비율로 정확도를 계산한다.

표 1. 컨텐츠 정보.
Table 1. contents information.

	$Cont_1$	$Cont_2$	$Cont_3$	$Cont_4$	$Cont_5$	$Cont_6$	$Cont_7$	$Cont_8$
The selected contents in the time slot t at day d, *Jan. 7, 2003*	0	0	0	1	0	0	0	0
The selected contents in the time slot $t+1$ at day $d-1$, *Jan. 6, 2003*	1	0	0	0	0	0	0	0
The really viewed content in the time slot $t+1$ at day d *Jan. 7, 2003* from test data	1	0	0	0	0	0	0	0

표 1은 선호도 전이 확률 계산을 위한 가중치 계산 및 정확도를 계산에 필요한 정보이다. 표 1에서 첫 번째 라인에서 보는 것과 같이 2003년1월7일 t 시간(7시~8시 사이)에 시청한 컨텐츠는 $cont_4$ 이다. 즉 현재 $cont_4$를 시청 중에 있다. 그리고 두 번째 라인과 같이 바로 전날 day $d-1$, $t+1$시간에 시청한 컨텐츠는 $cont_1$ 이다. 그리고 세 번째 라인은 정확도를 계산하기 위한 실험 데이터로 2003년1월7일 $t+1$시간(8시~9시 사이)에 실제 시청한 컨텐츠는 $cont_1$ 이다.

표 2는 2002년12월7일부터 2003년1월6일까지의 usage history로부터 시간 구간 t에서 $t+1$로 이동된 컨텐츠의 통계적 선호도 전이 행렬이다. 표 2에서 밝게 표시된 행은 현재 시청하고 있는 컨텐츠에 대한 선호도 전이 정보로 T(4,j)로 표시한다.

표 2. 선호도 전이 행렬.
Table 2. preference transition frequency matrix.

		Frequency of each contents selected in the time slot $t+1$ at day d from Dec. 7, 2002 to Jan. 6, 2003							
		$Cont_1$	$Cont_2$	$Cont_3$	$Cont_4$	$Cont_5$	$Cont_6$	$Cont_7$	$Cont_8$
Frequency of each contents selected in the time slot t at day d from Dec. 7, 2002 to Jan. 6, 2003	Cont1	3	0	2	3	1	0	0	0
	Cont2	4	0	3	3	1	0	0	0
	Cont3	2	0	3	2	0	0	0	0
	Cont4	4	0	4	5	0	0	0	0
	Cont5	0	0	1	1	0	0	0	0
	Cont6	0	0	0	0	0	1	0	0
	Cont7	0	0	0	0	0	0	2	0
	Cont8	0	0	0	0	0	0	0	0
T(4,j)		4	0	4	5	0	0	0	0

표 3에서 첫 번째 행은 훈련 데이터로부터 1개월간 d-1 (월요일) day, t+1 시간에 시청한 $cont_1$의 시청 빈도수를 나타낸다. 표 3에서 두 번째 행은 2일 연속해서 같은 컨텐츠를 소비한 컨텐츠의 빈도수에 가중치를 식(3)을 이용하여 계산한 결과이다. 세 번째 행부터 마지막 행까지는 시간 구간 t에서 시청한 $cont_4$로부터 시간 구간 t+1에서 시청 가능한 모든 컨텐츠들에 대한 선호도 전이 확률을 식(2)를 이용하여 계산하는 일련의 과정을 보여주고 있다. 표 3 마지막 행의 확률 값에 의해 2003년 1월7일 오후7~8시 사이에 $cont_4$를 시청하고 있는 상황에서 오후 8시~9시 사이에 추천하게 될 선호 컨텐츠는 확률 값이 가장 높은 $cont_1$을 추천하게 될 것이다. 표 1의 마지막 행에서 보듯이 추천된 $cont_1$은 실제로 고객에 의해서 시청된 컨텐츠임을 확인할 수 있다.

표 3. 컨텐츠 추천 예.

Table 3. Example of a recommendation contents.

		Frequency of each contents selected in the time slot t+1 at day d from Dec.7,2002 to Jan.6, 2003								$N_w(t+1)$
		$Cont_1$	$Cont_2$	$Cont_3$	$Cont_4$	$Cont_5$	$Cont_6$	$Cont_7$	$Cont_8$	
Frequency of Cont1 selected in the time slot t+1 at day d-1 from Dec. 7, 2002 to Jan.6, 2003	$Cont_1$	2	0	1	1	0	0	0	0	4
$W_{d-1,d}(I,j)$		1.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
$T(4,j)$		4	0	4	5	0	0	0	0	
$W_{d-1,d}(I,j)\,T(4,j)$		6.00	0.00	4.00	5.00	0.00	0.00	0.00	0.00	15.00
$\frac{W_{d-1,d}(1,j)T(4,j)}{\sum_{j=1}^{n} W_{d-1,d}(1,j)T(4,j)}$		0.40	0.00	0.266	0.333	0.00	0.00	0.00	0.00	

4. 시스템 구현 및 실험

4.1 구현 환경

본 논문에서 제안하는 선호도 전이 확률을 이용한 멀티미디어 컨텐츠 추천 시스템은 Java 2 SDK 와 J2ME Wireless Toolkit 환경에서 MIDlet을 이용하여 구현하였다. J2ME Wireless Toolkit은 MIDlet 응용 프로그램을 개발하기 위한 통합 개발도구이다.

그림 7은 본 논문에서 제안하고 있는 선호도 전이 확률을 이용하여 멀티미디어 컨텐츠를 추천하기 위한 프로토타입 시스템을 구현하여 실행하기 위한 WIPI(Wireless Internet Platform for Interoperability) 플랫폼 구조이다. WIPI는 모바일 단말기에 탑재되어 무선 인터넷을 통해 다운로드 된 응용 프로그램을 수행할 수 있는 환경을 제공하는 표준 플랫폼 규격이다. WIPI 2.0.1이 2004년 11월에 KWISF(Korea Wireless Internet Standardization Forum)[8]에 의해 채택되었다.

그림 7에서 HAL(Handset Adaptation Layer)은 단말기 하드웨어 독립성을 지원하기 위한 추상 계층이다. WIPI 응용 관리자(WIPI Application Manager)는 응용 프로그램 다운로드, 설치, 삭제, 정지 기능과 API(Application Programming Interface) 추가 갱신 기능을 제공한다. 기본 API는 WIPI 응용 프로그램 개발자를 위한 C 및 Java AP, J2ME(Java 2 Micro Edition) CLDC(Connected Limited Device Configuration/MIDP(Mobile Information Device Profile)로 구성된다. WIPI-C를 이용하여 구현된 응용 프로그램을 Clet이라 하며, WIPI-Java를 이용하여 구현된 응용 프로그램을 Jlet이라 한다. 그리고 J2ME CLDC/MIDP를 기반의 응용 프로그램을 MIDlet이라 한다[21].

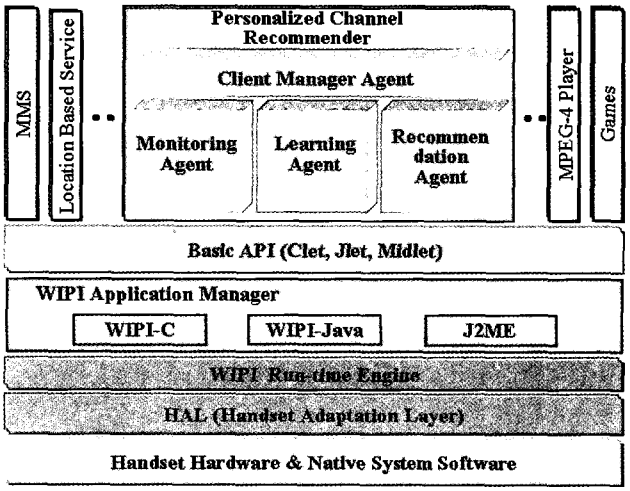

그림 7. WIPI 플랫폼에서 멀티미디어 컨텐츠 추천시스템 구조도.

Fig. 7. Schematic structure of the multimedia contents recommender implemented on WIPI platform.

4.2 구현 결과

그림 8은 J2ME 툴킷 에뮬레이터를 이용한 컨텐츠 추천 시스템 실행 화면이다. 고객이 특정 프로그램을 시청하는 동안 개인화 컨텐츠 추천 버튼을 선택하면 에뮬레이터 화면 오른쪽 상단에 개인화 컨텐츠 정보를 보여준다. 예를 들어 News 프로그램을 시청하는 동안 새로운 컨텐츠를 시청하기 위해 시청자가 개인화 컨텐츠 추천 버튼을 선택하면 개인화 컨텐츠 추천 정보가 화면 오른쪽 위쪽에 팝업 창으로 나타난다. 이렇게 함으로써 자신이 선호하는 컨텐츠를 찾는데 걸리는 시간을 절약할 수 있다.

그림 8. J2ME Toolkit 에뮬레이터를 이용한 추천 시스템 실행 화면.

Fig. 8. An example displaying the recommended contents on J2ME Toolkit Emulator.

4.3 성능 평가

본 논문에서 제안한 선호도 전이 확률 모델을 실험하기 위한 실험 데이터로는 한국의 대표적인 시장 조사 기관 중 하나인 AC Nielsen Korea에 위해 2002년 12월 1일부터 2003년 5월 31일까지 2,518명의 TV 시청자로부터 수집된 3,199,990건의 TV 시청 데이터를 이용하였다. TV 시청 데이터는 각 고객의 가정에 설치된 Set-Top Box를 이용하여 로그 인, 로그아웃 시간, 방송 시간과 요일, 시청 프로그램의 장르 등을 수집하였다. 실험 데이터가 가지는 각 프로그램은 8개의 장르, 즉 News, Entertainment, Drama & Movie, Information, Sports, Education, Children, Others로 구분한다. 실험 데이터로 이용된 전체 6개월 데이터 중 처음 5개월 데이터는 훈련 데이터로 이용하고, 나머지 1개월 데이터는 테스트 데이터로 이용하였다. 실험 데이터는 오후 7시부터 11시까지 시청한 데이터를 1시간 간격으로 구분하였으며, 시청 빈도수가 낮은 시간대의 데이터는 제외하였다. 실험 데이터에 대한 시청자의 선호 장르를 예측하기 위한 훈련 데이터는 길이의 일관성을 위해 시청자의 TV 시청 데이터 중 가장 오래된 날의 데이터는 제거하고 대신에 최근에 실험된 날의 데이터가 훈련 데이터에 포함된다. 성능 평가는 TOP-N 방법[22]과 본 논문에서 제안하는 선호도 전이 확률 모델의 정확도(precision)를 비교하였다. TOP-N 방법은 훈련 데이터에서 추천하고자 하는 특정 요일, 시간대에 시청한 정보를 이용하여 8개의 장르 중 시청 빈도수가 높은 순으로 N개를 추천하는 방법이다. 본 논문에서는 1시간에 시청하는 평균 컨텐츠 수를 고려하여 상위 2개를 추천하였다. 정확도는 다음과 같이 계산할 수 있다.

$$\text{정확도} = (\text{선호 컨텐츠 set} \cap \text{추천 컨텐츠 set}) * 100 / \text{추천 컨텐츠 set}$$

그림 9는 훈련 데이터의 수집 기간 변화에 따른 평가 모델간의 성능 변화를 알아보기 위한 그래프이다. 두 모델에 대한 정확도는 훈련 데이터의 크기가 1개월인 경우 비슷하지만, 4개월 이상인 경우 제안하는 모델의 추천 정확도가 70% 이상의 추천 정확도를 보인다. 이는 본 논문에서 제안하는 선호도 전이 확률 모델은 4개월 정도의 훈련 데이터를 이용하여 추천하는 것이 효과적임을 알 수 있다. 이는 연속적으로 소비한 컨텐츠의 빈도가 높을수록 제한하는 방법의 정확도가 높기 때문이다.

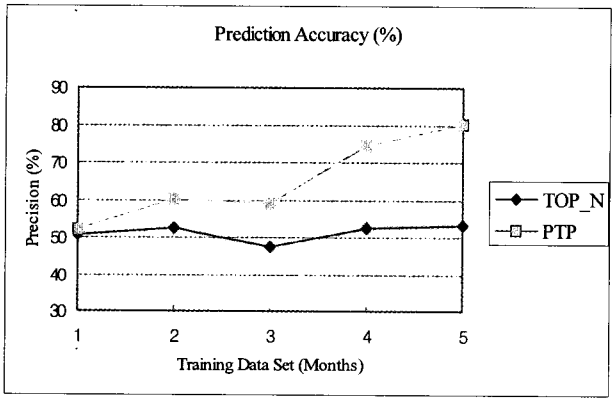

그림 9. 훈련 데이터 기간 변화에 따른 평가 모델간의 성능 비교.

Fig. 9. compare performances of the two methods by varying the size of the training data.

5. 결 론

본 논문에서는 개인화된 멀티미디어 컨텐츠를 추천하기 위한 방법으로 모바일 장치와 같은 클라이언트 쪽에서 선호도 전이 확률을 이용한 멀티미디어 컨텐츠 추천 시스템을 제안하였다. 제안하는 시스템은 클라이언트 관리자 에이전트, 모니터링 에이전트, 러닝 에이전트, 그리고 추천 에이전트를 포함한다. 추천 에이전트에서 고객의 컨텐츠 소비 성향에 대한 선호도 전이 확률을 계산하기 위한 추천 알고리즘을 제안하였다. 또한 실험 부분에서는 실제로 고객이 이용했던 usage history를 이용하여 본 논문에서 제안한 추천 알고리즘을 구현하고 실험하였다. 서버 쪽에서는 개인화를 제공하려면 많은 양의 데이터를 가져야 하기 때문에 클라이언트 쪽에서 제공되는 개인화 컨텐츠 추천 시스템이 서버 쪽에서 제공되는 개인화 추천 시스템에 비해 많이 가볍다. 제안하는 멀티미디어 컨텐츠 추천 시스템은 고객 개인의 usage history 데이터만을 이용하기 때문에 비교적 적은 양의 데이터를 취급한다. 따라서 모바일 디바이스가 가지는 제약 사항에 잘 적응할 수 있어 모바일 환경에서 실용화 될 수 있다. 또한 약 2000명의 TV 시청자가 시청한 데이터를 이용하여 본 논문에서 제안하는 선호도 전이 확률 모델의 성능이 전형적인 방법에 비해 정확도가 높음을 보임으로써 실질적으로 유용할 수 있음이 입증되었다.

참 고 문 헌

[1] Mobasher, B., Cooley, R., Srivastava, J., "Automatic Personalization Based on Web Usage Mining", Comm. of the ACM, Vol 43, 8, Aug. 2000.

[2] Maurice D. Mulvenna et al., "Personalization on the Net using Web Mining", Comm. of the ACM Vol. 43, 8, Aug. 2000.

[3] Myra Spiliopoulou, "Web Usage Mining for Web Site Evaluation", Comm. of the ACM Vol. 43, 8, Aug. 2000.

[4] Ibrahim Cingil et al., "A Broader Approach to Personalization", Comm. of the ACM Vol. 43, 8, Aug. 2000.

[5] Mike Perkowitz and Oren Etzioni, "Adaptive Web Sites", Communications of the ACM Vol. 43, 8, Aug. 2000.

[6] Udi Manber et al., "Experience with Personalization on Yahoo!", Comm. of the ACM Vol. 43, 8, Aug. 2000.

[7] Ee-Peng Lim, Wee-Keong Ng, "An Overview of the Agent-Based Electronic Commerce System (ABECOS) Project", Bulletin of the Technical Committee on Data Engineering, Vol. 23, No. 1, Mar. 2000.

[8] KWISFS.K-05-001, http://www.kwisforum.org/

[9] P. Resnick and H.R. Varian, "Recommender Systems," Communications of the ACM, Vol. 40, No. 3 , Mar. 1997.

[10] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer, 2000.

[11] P. Cotter and B. Smyth, "A Personalized Television Listing Service," Communications of the ACM, Vol. 43, No. 8, Aug. 2000.
[12] http://www.ptv.ie/
[13] 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.
[14] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proceedings of ACM Conference on Computer Supported Cooperative Work, 1994.
[15] 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.
[16] S. Kang, J. Lim, and M. Kim, "Modeling the User Preference on Broadcasting Contents Using Bayesian Networks," Journal of Electronic Imaging, to be appears on July, 2005.
[17] S. Kang, J. Lim, and M. Kim, "Statistical Inference Method of User Preference on Broadcasting Content," LNCS, Vol. 3514, May 2005.
[18] S. Kang, J. Lim, and M. Kim, "Modeling the User Preference on Broadcasting Contents Using Bayesian Belief Network Presentation," VCIP, Vol. 5308, Jan. 2002.
[19] J. Lim, S. Kang, and M. Kim, "User Preference Based Information Personalization for Easy Access for Multimedia Contents," WIAMIS, CD, Apr. 2004.
[20] 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/UM01/TV.html/
[21] http://www.java.sun.com/products/
[22] Mukund D., George K., "Item-Based Top-N Recommendation Algorithms", ACM Transactions on Information System, Vol. TBD, TBD 20 TBD, 2004.
[23] Athanasios Papoulis. Probability, Random Variables, and Stochastic Processes. McGraw Hill, 1991.

저 자 소 개

박성준(Sungjoon Park)
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

관심분야 : 개인화, 멀티미디어, 데이터마이닝, 모바일 정보시스템, 전자상거래
Phone : 041-850-9149
E-mail : sjpark@kcac.ac.kr

강상길(Sanggil Kang)
1989년 : 성균관대학교 전기공학과 학사
1995년 : Columbia University, 석사
2002년 : Syracuse University 박사
2004년~현재 : 수원대학교 컴퓨터학과 전임강사

관심분야 : 멀티미디어, 개인화, 유비쿼터스, 인공지능, 데이터마이닝
Phone : 031-229-8217
E-mail : sgkang@suwon.ac.kr

김영국(Kim, Young-Kuk)
1985년 : 서울대학교 계산통계학과 학사
1987년 : 서울대학교 계산통계학과 석사
1995년 : 버지니아대 컴퓨터과학과 박사
1995년~1996년 : 핀란드 VTT, 노르웨이 SINTEF DELAB 방문연구원
1996년~현재 : 충남대학교 전기정보통신공학부 부교수
2002년 8월~2003년 7월 : UC Davis 방문교수

관심분야 : 실시간 데이터베이스, 모바일정보시스템, 전자상거래시스템
Phone : 042-821-5450
E-mail : ykim@cnu.ac.kr