추천시스템의 성능 향상을 위한 시간스키마 적용 2단계 클러스터링 기법

Two-step Clustering Method Using Time Schema for Performance Improvement in Recommender Systems

부종수(Jong-Su Bu)*, 홍종규(Hong Jong-Kyu)**, 박원익(Park Won-Ik)**,
김 룡(Kim Ryong)*, 김영국(Young-Kuk, Kim)***

초 록

디지털 TV 채널 및 인터넷 상에서의 멀티미디어 컨텐츠의 홍수로 인해 사용자는 종종 자신이 선호하는 컨텐츠를 찾는데 어려움을 갖고 있으며, 또한 컨텐츠를 찾기 위해 많은 시간을 들이고 있다. 심지어 컨텐츠를 검색하는 동안 원하는 정보를 잃어버리는 경우도 있다.

고객들이 선호하는 컨텐츠를 추천하는 기존 시스템들이 가지는 문제점으로 사용자 수가 증가함에 따라 추천시간이 증가하는 확장성 문제와 새로운 고객의 경우 상품에 대한 선호도 정보가 부족할 경우 추천 정확도가 저하되는 희박성 문제가 있다.

본 논문에서는 이러한 문제점들을 해결하기 위해 협력적 필터링 방식에 기반한 2단계 클러스터링 기법을 제안한다. 1단계에서는 고객의 성과 나이와 같은 기본적인 사용자 정보만을 사용하여 추천하고, 2단계에서는 사용자의 동적인 성향 변화를 반영하기 위해 시간스키마를 적용하여 추천한다. 이렇게 추천된 결과의 피드백을 이용함으로써 계산시간의 단축과 예측정확도를 높일 수 있다.

ABSTRACT

With the flood of multimedia contents over the digital TV channels, the internet, and etc., users sometimes have a difficulty in finding their preferred contents, spend heavy surfing time to find them, and are even very likely to miss them while searching.

In this paper we suggests two-step clustering technique using time schema on how the system can recommend the user's preferred contents based on the collaborative filtering that has been proved to be successful when new users appeared.

This method maps and recommends users' profile according to the gender and age at the first step, and then recommends a probabilistic item clustering customers who choose the same item at the same time based on time schema at the second stage. In addition, this has improved the accuracy of predictions in recommendation and the efficiency in time calculation by reflecting feedbacks of the result of the recommender engine and dynamically update customers' preference.

키워드 : 개인화, 추천시스템, 2단계 클러스터링, 협력적 필터링
Personalization, Recommender System, Two-step Clustering, Collaborative filtering

---

본 논문은 한국산업기술평가원이 지정한 지역협력연구센터(RRC)인 충남대학교 소프트웨어연구센터의 지원으로 수행된 과제의 결과입니다.

* 충남대학교 차세대이동통신및서비스인력양성 누리사업단 조교
** 충남대학교 컴퓨터공학과 석사과정
*** 충남대학교 전기정보통신공학부 교수

1. 서 론

최근 세계적 인터넷 쇼핑몰인 Amazon.com을 비롯한 유수의 전자상거래 업체가 도입하고 있는 추천시스템은 고객이 경험하지 못한 상품이나 컨텐츠에 대해서 유사한 특성을 가지는 고객들의 반응을 토대로 만족할만한 상품을 추천해주는 시스템이라고 볼 수 있다. 따라서 고객이 어떤 상품에 관심이 있는지 또는 어떤 정보가 유용한지를 예측하는 일에 초점을 맞추며 이런 예측은 각 개인의 프로파일에 기반하여 개인화된다.

추천시스템 구현 기법으로는 초기에 사용자의 과거행위나 명시적 정보에 의해 미리 생성되어진 규칙을 가지고 적용하는 규칙기반 필터링 기법이나 다른 사람의 평가와는 무관하게 자신이 이미 평가한 아이템이나 컨텐츠를 기반으로 추천하는 내용기반 필터링 기법을 이용하였으나 최근에는 대상고객(target user)과 취향이 유사한 고객을 찾아서 그 고객이 좋아하는 아이템을 추천해주는 협력적 필터링 기법을 많이 사용하고 있다.

기존 추천시스템들이 가지는 대부분의 문제점은 사용자 수가 증가함에 따라 추천시간이 증가하는 확장성 (scalability) 문제와 새로운 고객의 경우와 같이 상품에 대한 선호도 정보가 부족할 경우 추천 정확도가 저하되는 희박성(saparsity) 문제가 있으며 이런 문제점들을 해결하기 위한 많은 연구와 실험이 이어져 왔으나 아직도 개선의 여지가 남아 있는 상황이다.[1] 선호도 예측의 문제점 중에는 어떤 상품을 추천할 것인가 하는 사용자 선호도 계산 시간에 관한 문제도 존재한다.

대표적인 협력적 필터링 기반 추천엔진은 미네소타 대학의 GroupLens 프로젝트에서부터 나온 것으로 피어슨 상관계수(Pearson correlation coefficient)를 이용하여 사용자간에 유사성을 구한다. 하나의 아이템에 대한 사용자의 선호도를 계산하기 위해서는 다른 모든 사용자와의 유사도를 계산하여야 하고 그 유사도를 바탕으로 또 다시 선호도 값을 계산하여야 한다. 이것은 사용자가 많은 전자상거래 시스템 내에서 실시간으로 행하기에는 너무 많은 연산 시간을 요구할 뿐만 아니라 예측 정확도면에서도 비효율적이다. 또한 처음 방문한 새로운 고객에게는 기본 프로파일 정보 이외에 어떤 선호경향도 파악할 수 없으므로 정확한 서비스를 제공할 수 없다.

본 논문에서는 고객의 기본 프로파일 정보 중 가장 변별력이 있는 성과 나이에 대한 1단계 그룹을 생성하고 그에 클러스터링 되도록 함으로써 그 집단의 선호 상품을 우선적으로 추천해주는 방식을 이용한다. 그리고 추천결과에 따른 피드백을 받아서 다시 시간 흐름에 따른 선호 경향별 2단계 그룹에 클러스터링하여 그 집단의 대표 선호 상품을 추천해주는 2단계 클러스터링 방법을 사용함으로써 새로운 고객에 대해 예측 정확도를 높일 수 있는 방법을 제안한다. 유사도가 높은 사용자들을 모아 시간에 따른 성향변화를 고려하고 2단계에 걸쳐 클러스터링하는 기법을 사용함으로써 선호도 계산시간의 단축과 더불어 동시에 예측정확도를 높이는 효과를 보이고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 협력적 필터링 기법에 관한 내용과 협력적 추천방식의 전형적인 모델인

GroupLens 프로젝트의 대표적인 연구에 대해 논하고 3장에서는 협력적 필터링 기반 추천 시스템에서 성능 향상을 위하여 제안한 시간 스키마를 이용한 2단계 클러스터링 기법에 대해 소개 한다. 4장에서는 설계 및 구현에 관한 부분으로 목표시스템 환경 구조와 각 구성 요소에 대해 기술하고 5장에서는 실험 데이터에 대한 내용과 제안방법에 대한 성능 평가를 수행한다. 6장에서는 결론 및 향후 연구 방향에 대해 기술한다.

2. 관련 연구

2.1 협력적 필터링 추천 알고리즘

2.1.1. 대표적인 두 가지 접근 방법

협력적 필터링의 경우 대개 두 부류의 알고리즘으로 나누어지는데, 하나는 메모리 기반 알고리즘(memory- based algorithms)이고 다른 하나는 모델 기반 알고리즘(model-based algorithms)이다.[1]

메모리 기반 알고리즘은 예측을 하기 위해 전체 사용자 데이터베이스를 관리한다. 다시 말하면 모든 아이템에 대한 모든 사용자의 선호도 데이터베이스를 유지하며, 전체 데이터베이스를 통해 계산을 수행한다. 수행된 결과를 통해 예측 값을 생성 한다.

모델 기반 알고리즘은 모델을 추정하거나 학습하기 위해 사용자 데이터베이스를 사용하며, 그렇게 생성된 모델은 예측을 위해 이용한다. 모델 중심 알고리즘은 사용자의 선호도를 사용자, 아이템, 그리고 평가 내용을 갖는 기술적인 모델로 재구성한다. 그리고 추천은 구성된 모델에 의뢰함으로써 이루어진다. 메모리 기반 알고리즘은 모델 기반 알고리즘보다 간단하며, 실제 상황에서 잘 작동하고 계속해서 새로운 데이터를 쉽게 추가할 수 있다. 그러나 단점은 데이터베이스의 크기가 커짐에 따라 처리 비용이 비싸진다는 것이다. 모델 기반 알고리즘의 경우, 일단 모델이 생성되면 예측은 빨리 계산되어진다. 그러나 데이터로부터 모델을 구축하는 것은 많은 시간이 소요되며 새로운 데이터를 추가하기 위해서는 모델을 전체적으로 재구성해야 하는 단점이 있다.

2.1.2. 클러스터링 모델

협력적 필터링기법에 기반을 둔 추천시스템이 웹 상에서 많이 이용되고 있다. 하지만 최근에 웹 사이트의 방문자 수와 이용 가능한 정보의 빠른 성장은 추천시스템에게 다음과 같은 문제를 야기 시켰다. 즉 추천시스템에게 질 높은 추천을 생성해야 하며, 짧은 시간 안에 수만 명의 사용자와 아이템에 대한 많은 추천을 수행해야 하고, 희박한 데이터로부터 높은 포괄성을 달성해야 하는 문제가 있다.[2]

추천시스템이 필요로 하는 새로운 기술로 제안된 클러스터링 기법은 유사한 선호도를 갖는 사용자를 식별하여 그룹으로 만든다. 일단 클러스터가 만들어지면 개인에 대한 예측은 그 클러스터 내에 있는 다른 사용자의 선호도 평균으로 이루어진다. 따라서 클러스터링 기법은 다른 방법보다 덜 개인화되기는 하지만 일단 클러스터가 완전히 만들어지면, 분

석해야 할 그룹의 수가 훨씬 적어져서 매우 우수한 성능을 발휘한다.

협력적 필터링의 확률모델에서는 조건부 독립 확률인 베이즈 분류기(Bayesian classifier)를 이용한다. 확률모델 방법에서는 선호도가 입력되지 않은 특정 클래스의 구성원이 주어져도 특정 아이템의 선호도를 예측할 수 있다. 조건부 분포에서 클래스와 선호도가 결합될 확률을 나타내는 확률 모델이 표준 나이브 베이즈(naive Bayes) 공식이다. 표준 나이브 베이즈 공식은 특정 클래스와 선호도 값이 완전한 사용자 집단이 관찰될 확률적 표현을 계산할 수 있다.[3]

따라서 본 논문에서는 이런 장점을 살려 초기 사용자 문제를 해결함에 있어 먼저 기본 프로파일 정보로 1단계 클러스터링을 수행하고 시점에 따라 같은 아이템을 선호하는 사용자를 대상으로 2단계 클러스터링을 동적으로 수행함으로써 다음 추천 시점에 선호경향을 동적으로 갱신하여 보여주고자 한다.

2.1.3. 이웃 선정 방법

협력적 필터링 방법은 사용자가 선정한 아이템에 대한 선호도를 계산할 때 모든 사용자들 예측 아이템에 평가한 값을 계산하게 된다. 이웃 선정 방법은 현재 사용자와 유사도가 비슷한 사용자들만을 이웃으로 선정하여 예측에 사용함으로써 예측 정확도를 높이는 방법이다. 이웃 선정 방법의 대표적으로 k-Nearest Neighbor 방법과 pre-Clustering 방법이 사용되고 있다. 이웃 선정 방법은 예측에 사용할 이웃의 수를 줄임으로써 예측 수행 시간의 단축 효과를 가져다준다.

k-Nearest Neighbor 방법은 예측을 수행할 사용자와 선호도가 가장 비슷한 임의의 k명을 구하고 예측에 사용하는 방법이다. 현재 사용자와 가장 높은 유사도를 갖는 사용자 k명을 선정하여 이웃으로 인정하고, 정해진 이웃을 이용하여 예측에 사용하게 된다. 협력적 필터링 방법은 전체 사용자 집합을 이웃으로 사용하기 때문에 선호도 성향이 반대인 사용자들도 예측에 사용되게 되지만, k-Nearest Neighbor 방법은 유사도가 가장 높은 k명만을 사용하기 때문에 선호도가 비슷한 사용자에 대한 정보만을 예측에 사용하게 된다. k-Nearest Neighbor 방법을 사용하여 이웃을 선정한 후, 예측을 수행하는 방법이 기본적인 협력적 필터링 방법보다 높은 예측 정확도를 보여주었다.[4]

pre-Clustering은 클러스터링 방법을 이웃 선정에 사용하는 방법이다. 클러스터링은 전체 사용자 집합을 유사도가 높은 여러 개의 클러스터로 나누는 방법으로 같은 클러스터 내의 사용자들 간의 유사도는 높고, 다른 클러스터에 포함된 사용자들의 유사도는 낮게 된다. 클러스터링을 통해 사용자 전체 집합을 여러 개의 클러스터로 나눈 후, 예측이나 추천에 사용되는 사용자가 포함된 클러스터를 찾는다. 이때 사용자가 포함된 클러스터 내의 사용자 집합이 현재 사용자의 이웃 후보가 되고, 이들 중에서 예측 아이템에 대해 선호도를 평가한 사용자들이 이웃으로 선정되고, 예측에 사용된다. 협력적 필터링 방법은 아이템에 대한 명시적인 선호도 점수를 사용하여 아이템에 대한 선호도를 예측하거나 추천하는 시스템이기 때문에 k-Means 클러스터링 알

고리즘이 pre-Clustering 에 주로 사용된다.[5]

본 논문에서는 1단계 클러스터링에서 성과 나이에 따른 pre-Clustering 기법으로 클러스터링을 수행한 후 그 안에서 k-Nearest Neighbor 방법으로 다시 클러스터링을 수행하는 방법을 사용하여 연산시간과 예측 정확도면에서 성능 향상을 가져올 수 있는 방법을 제안한다.

2.2 GroupLens 프로젝트

2.2.1. GroupLens 프로젝트

미네소타 대학의 GroupLens는 같은 취향이나 취미를 가진 사람들의 정보를 이용해 추천할 때 도움을 주는 시스템으로 어떤 사람의 아이템에 대한 관심도를 예측하기 위하여 다른 고객들의 평가를 모아 이용하는 분산 시스템이면서, 일반적인 정보에 적합하도록 만들어진 필터링 기술이다. GroupLens는 인터넷 뉴스를 추천하는 시스템으로 소개된 이래 아마존, 리바이스, CDNOW 등과 같은 사이트들에서 여러 형태로 널리 사용되고 있다.[6]

GroupLens는 두 가지 평가 방법을 이용한다. 첫째 대상고객과 어떤 고객의 평가가 가장 유사한지 연관성을 계산하는 것이고, 둘째 그 유사한 고객의 평가를 근간으로 새로운 아이템에 대한 평가를 예측하는 방법이다. 여기서 피어슨의 [-1,1]의 값을 갖는 상관계수를 이용하여 상관관계를 구하는 방법은 [식 1]과 같다. 값에 따라 1로 접근할수록 양의 상관관계(대상고객과 다른 회원은 유사한 성향을 가진다)를 가지고 -1로 접근할수록 음의 상관관계를 가지며, 0으로 접근할수록 서로간의 상관관계가 없다는 의미로 해석된다.

통계적 협력적 필터링의 일반적인 공식화는 GroupLens 프로젝트에서 처음으로 이루어졌다. 가중치에 대한 기본 공식으로 피어슨 상관계수를 정의하였는데 이용자 a와 i의 상관관계는 다음과 같다.

$$u(a,i) = \frac{\sum_j (v_{a,j} - \overline{v_a})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_j (v_{a,j} - \overline{v_a})^2 \sum_j (v_{i,j} - \overline{v_i})^2}}$$

[식 1] 피어슨 유사관계 이용한 두 사용자의 상관관계

이 식에서 j는 사용자 a와 i가 동시에 선호도를 입력한 아이템의 개수이다. 또한 가중치 w (a ,i)는 -1에서 1까지의 값을 가지며, 음수 값은 부정적인 상관관계를 나타내며 양수 값은 긍정적인 상관관계를 나타낸다.

2.2.2. GroupLens의 대표적인 연구

GroupLens 프로젝트에서 협력적 추천시스템의 단점을 보완하기 위한 연구가 활발히 진행 중이며 정확도의 개선을 위해 유사도 가중치, 분산 가중치 등의 변형된 유사도 계산 공식을 제안했다.

Shardanand는 평균을 이용하는 방법과 피어슨 관계 계수, 제한적 피어슨 관계 계수, 상품-상품 유사도를 이용하여 사용자끼리의 유사도를 측정하고 사용자 유사도에 한계치를 주어 그에 따른 결과를 연구하였다. Herlocker et al의 논문[7]에서는 정확도의 개선을 위해 공통으로 평가한 항목 수를 유사도 계산에 반영시키는 중요도 가중치(significance weighting) 변형된 유사도 계산 공식을 제안했다. Breese

의 논문[8]에서는 피어슨 관계 계수와 벡터 유사도 같은 Memory-based 기법을 사용하고 그들 각각에 기본 평가 값을 사용하여 정확도와 coverage를 향상시키는 것에 관한 연구와 기존의 확률적인 방식인 Bayesian 방식의 Memory-based 기법을 이용한 방식을 응용하는 방법에 관한 연구를 하였다.

계산량을 줄이고자 하는 시도로서는 사용자의 유사도를 미리 계산하여 유사도가 높은 사용자를 선정하여 두었다가 선호도 값을 계산할 당시에는, 유사도가 높았던 일부 사용자와의 유사도만을 계산하여 계산 시간을 단축한다. 이러한 유사도가 높은 사용자의 집단을 neighborhood라 한다. 이러한 neighbor의 선정 기준을 크게 일정 수준의 유사도를 가진 사용자를 선정하는 임계값(threshold) 방식과 유사도 상위 N명을 선정하는 방식이 있다. 이러한 neighbor의 선정은 계산 시간의 단축과 더불어 유사 사용자만을 계산에 사용하기 때문에 선호도 예측 값 계산의 정확도를 개선한다.

Coverage를 높이는 방법으로는 앞에서 설명한 기본값을 이용하는 방식과 함께 항목을 분류하여 계층구조를 구성하고 사용자 사이에 공통항목이 없더라도 상위 레벨의 같은 카테고리에 속한다면 두 사용자간의 유사도를 계산할 수 있는 방법이 제시되었다.

본 논문은 이러한 대표적 연구들을 바탕으로 그동안 관심이 적었던 협력적 필터링의 중요 부분에 대해 서술 한다. 또한 본 논문은 2단계 클러스터링 기법을 적용한 새로운 사용자 문제를 해결하는 방법과 시간스키마 적용하여 예측 정확도와 계산 시간에서의 시스템 성능 개선 방법을 제안 한다.

3. 문제 해결 방안

3.1. 새로운 고객 문제

3.1.1. 기존 방식의 문제점

처음 방문한 새로운 고객에게는 그 고객의 프로파일 정보 이외에 어떤 선호경향도 파악할 수 없으므로 정확한 서비스를 제공할 수 없다. 따라서 기존의 협력적 필터링 시스템은 예측하기 전에 전형적으로 새로운 고객이 초기 정보 수집단계에서 요청 항목에 대해 평가하도록 요구한다. 그리고 그 평가를 기반으로 코사인 함수나 피어슨 상관계수식을 적용하여 사용자간의 유사도를 구한 다음, 아이템에 대한 선호도를 예측하는 방식을 사용한다. 그러나 모든 사용자간의 유사도 계산을 기본으로 하는 방식은 데이터의 희박성(sparseness)으로 인해 실제로 적용하는데 다음과 같은 중요한 문제점을 가지고 있다.

첫째, 고객-아이템 행렬로 구성된 고객 프로파일에 선호도를 부여한 아이템 수가 적을수록 유사도를 정확하게 계산할 수가 없어서 선호도 예측의 정확도면에서 비효율적이다. 특히 아이템에 대한 평가가 거의 없는 새로운 고객일 경우에는 더욱 더 중요한 문제가 된다.

둘째, 유사도가 낮은 사용자들이 많을수록 실제 평가 값이 극단적으로 갈 때 선호도 예측 값의 오차가 커진다.

셋째, 모든 사용자의 평가 데이터 전체를 입력 데이터로 사용함으로써, 상호관계가 없는 사용자간의 유사도 계산으로 인해 불필요한 시간이 소요된다.

3.1.2. 2단계 클러스터링을 이용한 제안방식

기존 방식의 문제점을 보완하여 시간에 따른 고객 성향을 반영하면서 예측정확도를 높일 수 있는 방법으로 제안한 동적인 2단계 클러스터링 기법의 기본 개념은 다음과 같다.

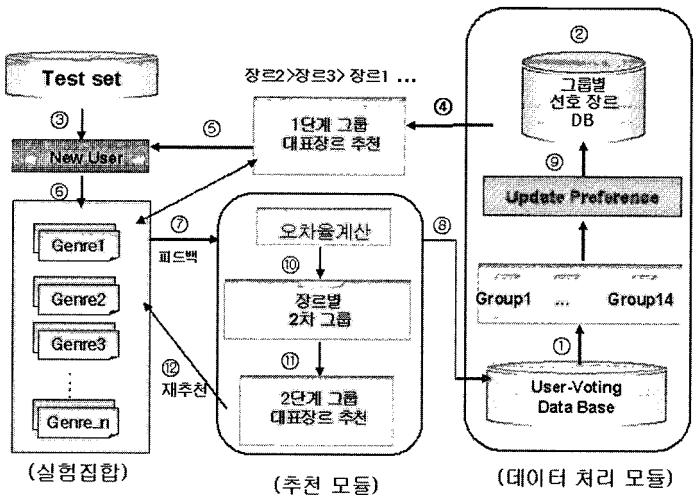
1단계 클러스터링 과정에서는 아이템에 대한 어떤 평가정보도 없는 새로운 고객에게 어느 정도 성향에 맞는 아이템을 제시하기 위하여 가장 변별력이 있는 성과 나이에 대한 1단계 클러스터링을 수행한다. 그리고 그룹 프로파일에서 높이 평가된 아이템 순으로 우선적으로 추천하는 것이다. 이렇게 하면 다른 사용자들과 차별화할 수 있고, 전체 사용자에 대한 유사도를 계산하지 않아도 되므로 연산시간도 줄일 수 있다.

2단계 클러스터링 과정에서는 추천 결과에 대한 사용자의 피드백을 받는다. 전 시점에서 같은 장르를 본 다른 사용자들을 대상으로 2단계 클러스터링을 동적으로 수행하여 현 시점에서 어떤 장르를 확률적으로 가장 선호 하였는지를 계산하여 추천해 준다. 이때 전 시점이라는 것은 과거 선호 영역으로 월을 기준으로 윈도우 사이즈(WS=1, 2, 3) 만큼의 기간을 의미한다. 이런 과정이 N번 반복 될수록 시점에 따른 최근 선호경향을 동적으로 반영할 수 있기 때문에 새로운 고객의 선호 아이템에 대한 예측 정확도가 개선될 뿐만 아니라 효율성면에서도 향상된 시스템을 구축할 수 있다.

3.1.3. 새로운 고객을 위한 시나리오

step1.

1. <그림 1>에서 새로운 고객이 VOD(Video On Demand) 사이트를 방문 한다.
2. 추천시스템은 일단 새로운 고객에 대한 기본 정보이외에 아이템에 대한 어떤 평가 정보도 없기 때문에 1단계 그룹 즉 10



<그림 1> 새로운 고객에 대한 추천 방식

대 남자 그룹에 클러스터하여 아이템을 추천한다.

가정1. 추천엔진에서 1단계 클러스터링 결과로 추천해주는 대표 장르의 추천 선호도 순서가 〈표 1〉과 같다고 가정하자.

〈표 1〉 1단계 클러스터링 결과 추천 선호도

1. 애니메이션
2. 드라마
3. 코미디_패밀리
4. 액션_패밀리
5. 액션_스릴러
6. 코미디
7. 패밀리
8. 애니메이션_드라마

가정2. 테스트 집합의 새로운 고객은 타임스탬프 순으로 했을 때 1.액션_패밀리, 2.코미디_패밀리, 3.코미디_패밀리, 4.드라마, 5.애니메이션 장르 순으로 영화를 보았다고 하자.

stcp2.

고객이 평가한 아이템에 대한 선호도 값과 추천엔진에서 추천된 아이템의 선호도 값을 비교하여 오차율을 계산한다. 오차율은 예측 평가 값과 실제 평가 값 사이의 평균 오차율 MAE(Mean Absolute Error) 방식을 이용하였다.

step3.

1. 피드백을 받은 시스템은 오차율 정보를 데이터베이스 에 반영한다.
2. 추천엔진은 전 시점에서 대상고객과 같은 장르를 선택한 고객들을 대상으로 2단계 클러스터링 을 수행한다.
3. 그리고 현 시점에서 2단계 클러스터링 안에 포함된 고객들을 대상으로 선호 아이템들의 확률을 구한다.
4. 2단계 그룹에서 가장 높은 확률을 보이는 아이템 순으로 대표 장르가 새로이 생성되고, 그룹 선호도 테이블과 사용자의 선호도 테이블이 갱신되어 다음 추천 장르를 순 서대로 보여준다.

따라서 step2와 step3을 반복 적용 할수록 시점에 따른 월별 시간 성향을 반영할 수 있기 때문에, 정적인 수집 데이터를 바탕으로 다른 모든 사용자와의 유사도를 계산하여 선호도 값을 예측하는 방법보다 정확도가 높은 추천을 해 줄 수 있다.

3.2. 시간적 윈도우 스키마

기존 논문에서 해결되지 않은 가장 큰 문제점 중의 하나가 시간에 따른 사용자의 선호 경향 파악이다. 시간적 흐름 성향을 반영하는 사전확률 방법을 이용하여 앞으로의 사용자 선호도를 표현하기 위해서는 많은 고객의 이용 히스토리가 필요하고 이에 따른 자료 수집 비용이 많이 들어간다는 문제점이 있다. 따라서 기존 논문에서는 보통 정적인 방법을 이용하여 전체 데이터에 대한 상관관계를 계산하는 방식을 이용함으로써, 시간적인 성향변화에 대한 분석이 고려되지 않았다.

시간적 스키마를 적용함에 있어 염두에 두어야 할 점은 무엇 보다 최근에 수집된 데이터가 향후 선호도 확률을 계산하는데 적절히 이용되지 못하고 있다는 점이다. 사용자의 과거이용 데이터의 크기가 새로운 데이터 크기에 비해 월등히 크기 때문에 주로 과거 데이

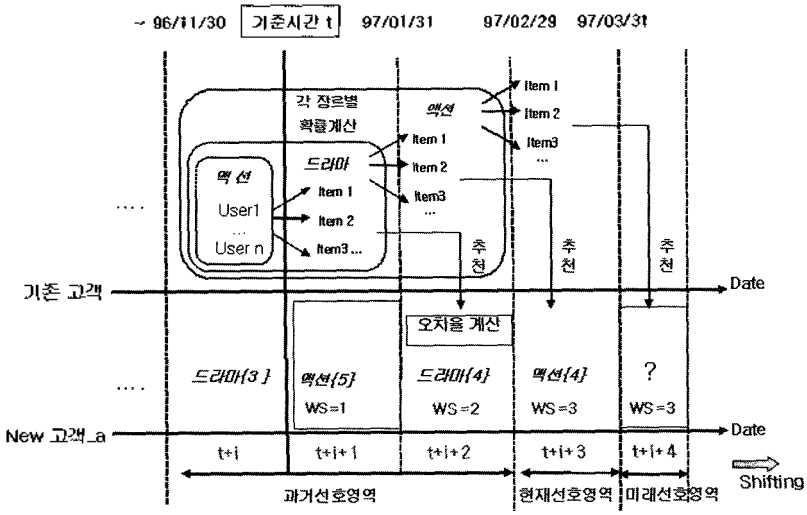
터에 의해 선호도가 좌우된다. 특히 선호도가 시간에 따라 자주 변하는 경향을 가지는 고객들에게는 과거 데이터 보다는 최근 데이터가 미래 선호도를 예측하는데 더 많은 영향을 줄 것이다.

본 논문에서는 시간 스키마 적용을 위해서 사용자들의 히스토리 데이터를 시간 축(월)을 기준으로 여러 개의 작은 집합으로 나눈 다음 2단계 클러스터링을 전 시점에서 같은 장르를 본 고객에 대해서만 윈도우 사이즈 안에서 동적으로 수행한다. 그리고 다음 추천시점에서 그 고객들이 평가한 장르들 중 대표 선호 장르를 순서대로 정렬하여 보여 준다. 즉 과거 평가 데이터를 이용하여 추천하는 것이 아니라 최근 윈도우 사이즈 만큼의 데이터를 이용하여 향후 선호 아이템을 예측하여 추천해 줌으로써 시간에 따른 고객 선호 경향 반영에 큰 역할을 한다.

〈그림 2〉는 윈도우 사이즈가 3인 시점에서의 선호 영역과 2단계 클러스터링을 수행하는 과정을 나타낸 그림이다.

윈도우 사이즈는 클러스터링 되는 영역 사이즈로 과거선호 영역이 된다. 새로운 고객이 평가하는 장르 패턴을 적용하여 시점별로 클러스터링 되는데 본 논문에서는 최대 3으로 하였으며, 3을 넘으면 오른쪽으로 이동(Shifting) 되어 계속 3을 유지한다. t는 기준시간으로 훈련 집합과 실험 집합을 나누는 기준점이 된다. t+[i]+n은 시간성향 반영을 위해 한달 간격으로 시간 축을 나누었다. 장르i{n} (예: 액션{5}) 에서 n은 월 간격으로 시간 축을 구분했기 때문에 한달 동안 장르 [i]를 평가한 회수가 된다.

과거선호영역은 현시점을 기준으로 윈도우 사이즈만큼의 전 평가 데이터 영역을 의미하고, 현재 선호영역은 평가하는 현재 시점을,



〈그림 2〉 시간에 따른 윈도우 스키마

미래선호 영역은 예측하고자 하는 바로 앞 시점의 선호 영역을 의미한다. t+[i]-2 시점에서의 새로운 고객이 평가한 대표 장르는 액션이고, t+[i]-1 시점에서는 드라마, t+[i] 시점에서는 액션이다.

3.3. 시간스키마를 이용한 2단계 클러스터링

새로운 고객의 미래 선호 영역에 대한 선호 장르 예측을 시간 스키마에 따라 2단계 클러스터링에 적용하는 과정은 다음과 같다.

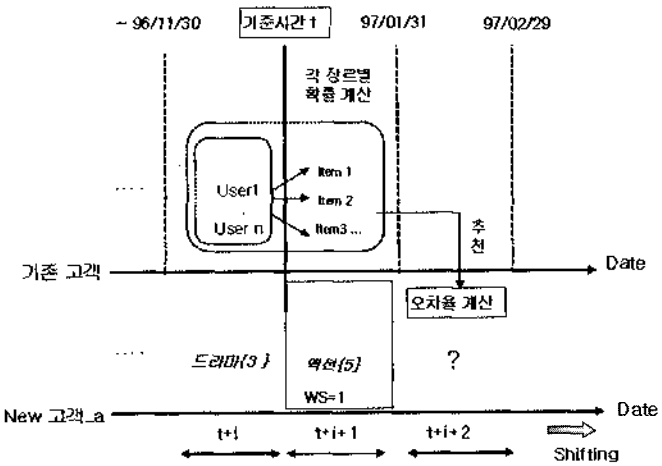
1) 윈도우 사이즈가 1인 경우(새로운 고객이 처음 방문한 경우)

〈그림 3〉에서 기준 시간이 t이므로 t+[i]+1 시점에 새로운 고객이 방문한다. 이때 윈도우 사이즈(WS: Window Size) 는 1이다.

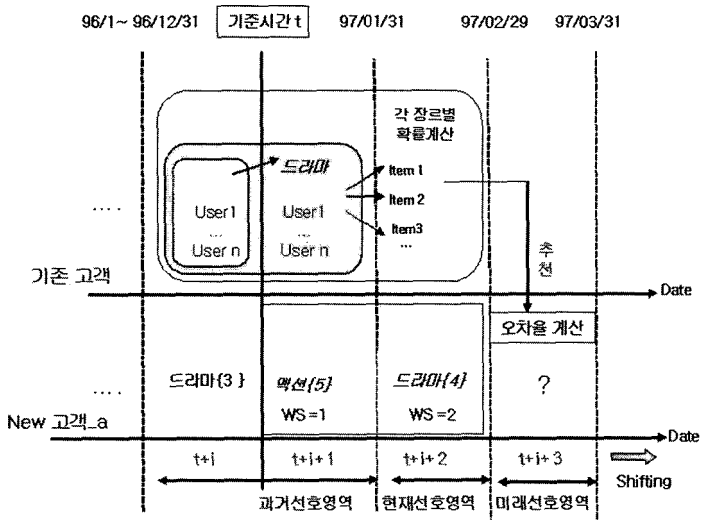
step1. t+[i] 시점에서 새로운 고객이 액션을 평가했으므로 추천엔진에서는 t+[i]-1시점(윈도우 사이즈 WS=1)에서 액션을 본 다른 고객을 대상으로 2단계 클러스터링이 동적으로 수행된다.

step2. t+[i]시점에서는 t+[i]-1 시점에서 클러스터링 된 고객을 대상으로 그들이 평가한 장르에 대한 평가점수와 빈도수를 적용하여 평균값이 높은 순서대로 가장 선호하는 장르를 추출하여 t+[i]+1 시점에 새로운 고객에게 추천한다. 즉 t+[i]-1 시점에서 액션을 본 다른 고객들이 t+[i]시점에 A라는 장르를 선호하였으므로 t+[i]시점에 액션을 선택한 새로운 고객도 t+i+1시점(미래 선호 영역)에 A라는 장르를 선호할 확률이 높다는 것이다.

step3. 추천엔진에서는 추천 장르 선호도 값과 고객이 평가한 장르 선호도 값의 차이로 오차율을 구하여 추천할 때 마다 예측 정확도를 평가한다. 그리고 새로운 고객의 선호 테이블을 갱신한다. 추천엔진에서의 오차율 계



〈그림 3〉 WS=1일때, 시간스키마 적용한 2단계 클러스터링

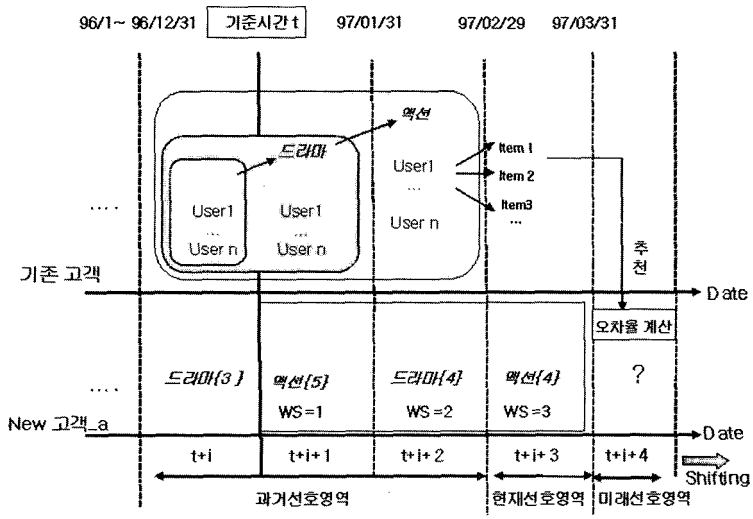


〈그림 4〉 WS=2일때, 시간스키마 적용한 2단계 클러스터링

산과 선호테이블 갱신은 WS=2, WS=3일때도 같은 과정을 거친다.

2) 윈도우 사이즈(WS) 가 2일 때

step4. 〈그림 4〉에서 t+[i]시점에서 고객이 드라마를 평가 했으므로 추천엔진에서는 t+[i]-2시점과 t+[i]-1시점(윈도우 사이즈



〈그림 5〉 WS=3일때, 시간스키마 적용한 2단계 클러스터링

WS=2)에서 액션을 보고 드라마를 본 다른 고객을 대상으로 2단계 클러스터링이 동적으로 수행된다.

step5. 클러스터링 된 고객을 대상으로 t+[i] 시점에서 확률적으로 가장 선호하는 장르를 추출하여 t+[i]+1시점에 추천한다.

3) 윈도우 사이즈(WS) 가 3일 때

step5. <그림 5>에서 t+[i]시점에서 고객이 액션을 평가했으므로 추천엔진에서는 t+[i]-3시점과 t+[i]-2시점 그리고 t+[i]-1시점(윈도우 사이즈 WS=3)에서 액션을 보고 드라마를 보고 액션을 본 다른 고객을 대상으로 2단계 클러스터링이 동적으로 수행된다.

step6. 클러스터링 된 고객을 대상으로 t+[i] 시점에서 확률적으로 가장 선호하는 장르를 추출하여 t+[i]+1시점에 추천한다.

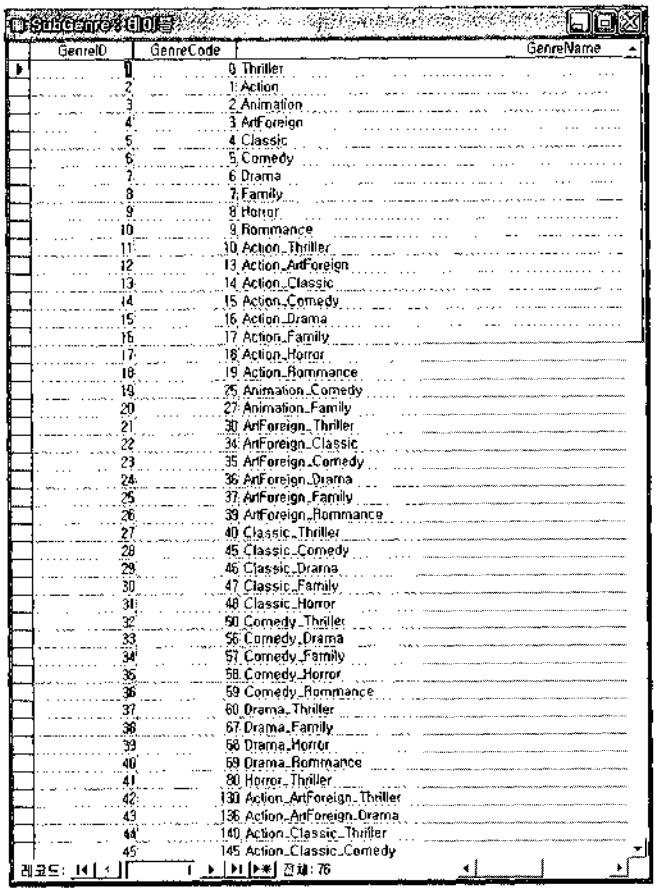
위와 같은 과정을 N번 반복 적용할수록 오른쪽으로 이동(shifting)되면서 윈도우 사이즈만큼 고객이 평가 패턴에 따라 2단계 클러스터링이 동적으로 수행된다. 윈도우 사이즈가 길다는 것은 과거 선호 경향을 많이 반영하는 결과가 되므로 최근 3개월 이내의 선호경향 반영을 위하여 윈도우 사이즈는 3으로 한다. 윈도우 사이즈가 3보다 커지면 오른쪽으로 한시점이 이동 되어 기준 시점에서 항상 윈도우 사이즈 3을 유지한다.

3.4. 장르 세분화에 의한 정보 제공

실험 데이터로 사용된 EachMovie 데이터는 크게 10개 장르 (액션, 애니메이션, 외국예술, 고전, 코미디, 드라마, 가족, 공포, 스릴러)로 구성되어 있지만 각 영화마다의 장르 속성은 그것들의 조합으로 이루어져 있다. 다시 말하면 The Story라는 영화가 있을 때 이 영화의 장르 속성은 애니메이션이면서 가족영화라는 특성을 가지며 Richard III 라는 영화는 외국예술 장르이면서 드라마 장르에 속한다. Natural Born Killers 라는 영화는 액션이면서 공포이면서 스릴러인 영화이다. 하지만 기존 연구에서는 이런 장르 특성을 무시하고 대표되는 장르만을 가지고 추천이 이루어졌기 때문에 보다 세밀한 방법으로 개인에 대한 선호 경향을 정확히 반영하지 못하는 결과를 초래했다.

하나의 영화가 하나의 장르에만 속한다는 극단적인 가정 하에 이루어지는 추천시스템일 경우, 추천 후보자수가 그만큼 많아지게 되므로 고객의 입장에서는 유용하지 못한 정보가 될 수도 있으며, 각 영화 특성이 정확히 반영 되지 못하는 점이 있기 때문에 추천 결과에 대한 만족도도 떨어질 수밖에 없다. 따라서 본 논문에서는 각 영화에 따른 장르 조합을 서브 장르로 분류하여 76개의 장르를 <그림 6>과 같이 추출하였으며, 그에 대한 사용자의 모든 선호도를 바탕으로 그룹별 대표 장르를 <그림 7>과 같이 추출하였다.

기존 Each movie 데이터를 이용한 추천시스템은 방문하는 사용자의 성향과 무관하게 장르가 정적으로 고정되어 보여 짐으로써 새로운 사용자가 들어 왔을 때도 정적인 구조를 갖는다는 문제점이 있다. 반면 시간 스키마를 적용한 개선된 VOD 추천시스템은 2단계 클러스터링이 월별 시점에 따라 동적으로 생성되기 때문에 방문하는 사용자의 성별, 나이

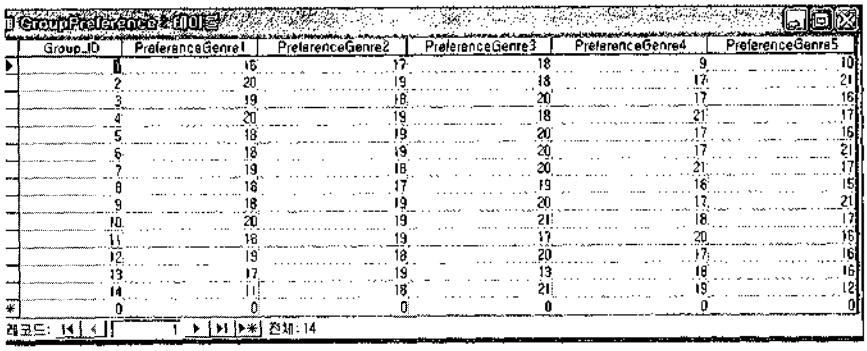


SubGenre 테이블

GenreID	GenreCode	GenreName
1	0	Thriller
2	1	Action
3	2	Animation
4	3	ArtForeign
5	4	Classic
6	5	Comedy
7	6	Drama
8	7	Family
9	8	Horror
10	9	Rommance
11	10	Action_Thriller
12	13	Action_ArtForeign
13	14	Action_Classic
14	15	Action_Comedy
15	16	Action_Drama
16	17	Action_Family
17	18	Action_Horror
18	19	Action_Rommance
19	25	Animation_Comedy
20	27	Animation_Family
21	30	ArtForeign_Thriller
22	34	ArtForeign_Classic
23	35	ArtForeign_Comedy
24	36	ArtForeign_Drama
25	37	ArtForeign_Family
26	39	ArtForeign_Rommance
27	40	Classic_Thriller
28	45	Classic_Comedy
29	46	Classic_Drama
30	47	Classic_Family
31	48	Classic_Horror
32	50	Comedy_Thriller
33	56	Comedy_Drama
34	57	Comedy_Family
35	58	Comedy_Horror
36	59	Comedy_Rommance
37	60	Drama_Thriller
38	67	Drama_Family
39	68	Drama_Horror
40	69	Drama_Rommance
41	80	Horror_Thriller
42	130	Action_ArtForeign_Thriller
43	136	Action_ArtForeign_Drama
44	140	Action_Classic_Thriller
45	145	Action_Classic_Comedy

레코드: 1 / 전체: 76

〈그림 6〉 대표 장르에서 추출된 서브 장르 정보

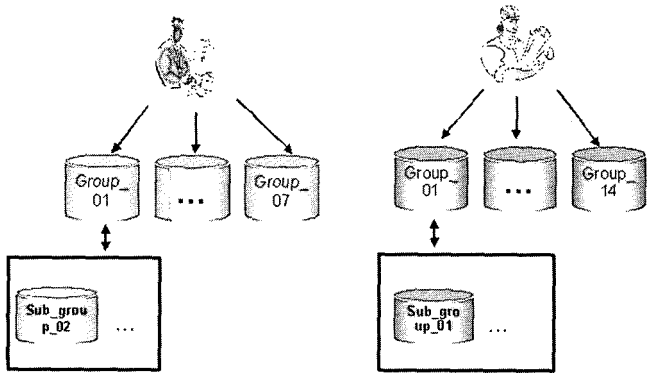


GroupPreference 테이블

Group_ID	PreferenceGenre1	PreferenceGenre2	PreferenceGenre3	PreferenceGenre4	PreferenceGenre5
1	16	17	18	9	10
2	20	19	18	17	21
3	19	18	20	17	16
4	20	19	18	21	17
5	18	19	20	17	16
6	18	19	20	17	21
7	19	18	20	21	17
8	16	17	19	16	15
9	18	19	20	17	21
10	20	19	21	18	17
11	18	19	17	20	16
12	19	18	20	17	16
13	17	19	13	18	16
14	11	18	21	19	12
0	0	0	0	0	0

레코드: 1 / 전체: 14

〈그림 7〉 성/나이 그룹별 선호도 장르 정보



〈그림 8〉 동적 2단계 클러스터링 모델

또는 선호 경향에 따라 76개의 장르가 동적으로 학습되면서 실시간적으로 갱신되어 보여진다. 이런 방법은 추천 아이템 분류가 적을 경우에는 별 상관이 없지만 많은 아이템 속에 선호 아이템을 찾고자하는 경우 가장 선호하는 아이템 순으로 사용자에 따라 다르게 보여짐으로써 사용자가 만족하는 아이템을 찾는 노력을 최소화할 수 있고 만족도가 높은 추천을 할 수 있다.

인터넷에서의 뉴스추천이라든가 디지털 TV와 같이 수많은 채널로 이루어져서 사용자들이 많이 컨텐츠 속에서 선호 컨텐츠를 찾기 어려운 경우 좋은 해결방안이 되리라 본다.

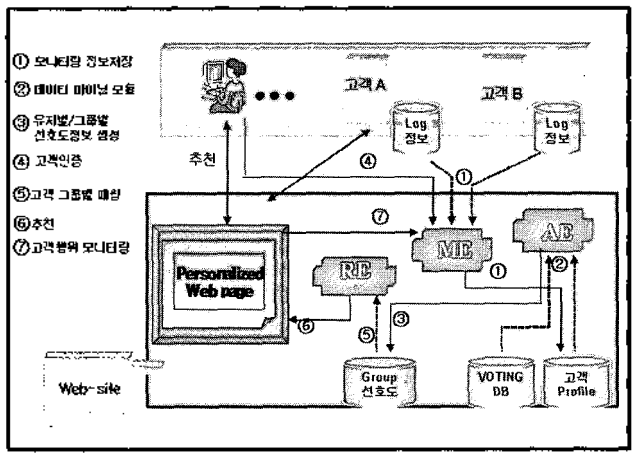
4. 목표 시스템 설계 및 구현

4.1. 목표 실험 시스템

고객의 추천 요구가 있을 때 평가 데이터 전체를 입력 데이터로 사용하여 고객이 좋아할 제품을 예측한다. 그러나 이러한 방법을 사용할 때 실제적으로 전자상거래 시스템에서 실시간으로 추천이 수행 될 경우 입력 데이터의 수가 폭발적으로 증가하므로 추천시스템의 수행 시간이 실시간으로 처리되기에 어려움이 있다.

따라서 본 논문에서 제안하는 추천시스템에서는 〈그림 8〉과 같이 1단계 클러스터링과 2단계 클러스터링 모두에서 유사도가 높은 사용자들을 모아 사용자의 수를 확연히 줄임과 동시에 시간 스키마에 따른 2단계 클러스터링을 수행함으로써 예측 정확도를 향상시킬 수 있는 개선된 추천시스템을 목표로 한다.

〈그림 9〉는 본 논문에서 제안한 2단계 클러스터링 기법을 적용할 목표 시스템 환경구조이다. 웹사이트의 Dynamic Learning VOD(Video on Demand)시스템은 고객이 원하는 비디오를 찾을 수 있도록 도와주는 추천 사이트로 고객들에게 비디오를 대여해주고, 그 대여 정보를 데이터베이스에 저장하며, 피드백을 줄 때마다 동적으로 선호도 정보를 갱신하여 최신 선호 경향을 반영하여 준다. 시스템 구성 엔진 모듈은 데이터 처리 모듈인

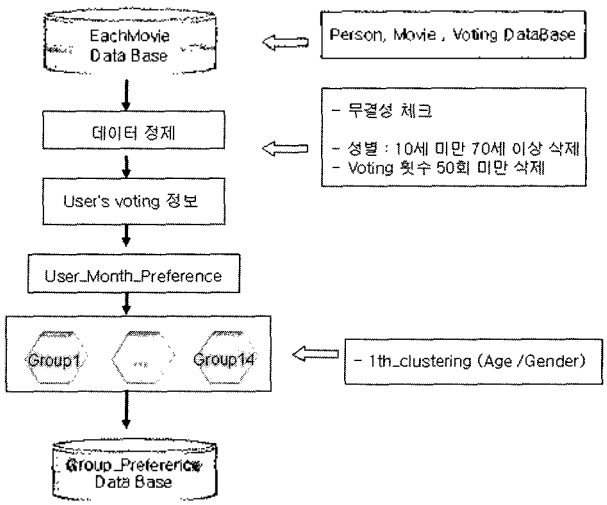


〈그림 9〉 VOD 시스템 아키텍처

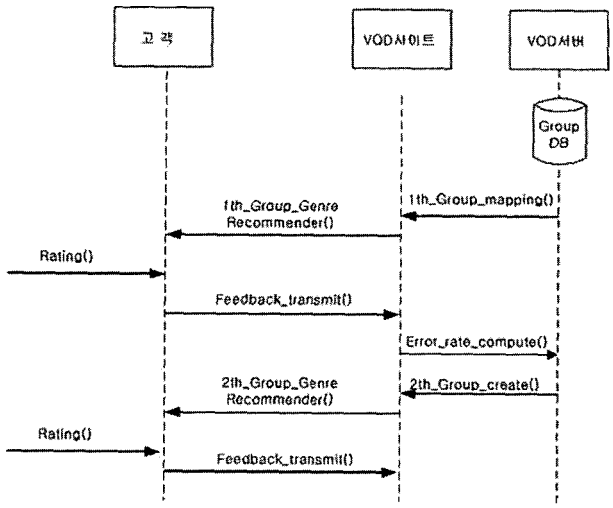
분석엔진(AE) 그리고 고객행위를 모니터링 하는 모니터링 엔진(ME) 핵심 모듈인 추천 엔진(RE)으로 나누어 볼 수 있다. 다음에서 각각의 기능에 대해 설명한다.

4.2. 분석 엔진(Analysis Engine)

분석엔진은 Off-Line 상에서 이루어지는 모듈로써 〈그림 10〉과 같이 데이터 정제과정을 거쳐 최적화시킨 후, 성과 나이에 따른 1단계



〈그림 10〉 분석엔진에서의 1단계 클러스터링

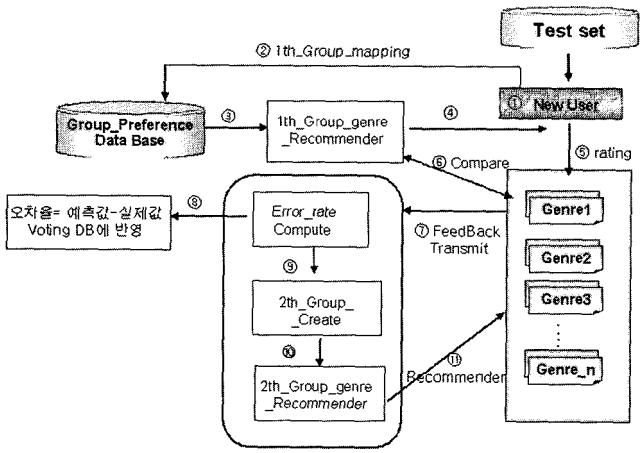


〈그림 11〉 모니터링엔진에서의 시퀀스 다이아그램

클러스터링을 수행하여 각 사용자별 그룹별 선호 장르 정보를 생성한다.

4.3. 모니터링 엔진(Monitering Engine)

〈그림 11〉은 새로운 고객이 VOD사이트에 방문하였을 경우, 시스템은 고객의 회원가입을 요청하고 고객이 프로파일 정보를 입력하면, VOD 서버에 회원으로 등록된다. 로그인한 후 인증절차를 거쳐 VOD 사이트의 서비스를 받을 수 있다.



〈그림 12〉 추천엔진에서의 2단계 클러스터링

4.4. 추천 엔진(Recommendation Engine)

추천엔진은 On-Line 상에서 이루어지는 핵심 엔진 모듈로써 기존의 협력적 필터링 기법의 틀을 유지하면서 시간 스키마를 적용한 2단계 클러스터링 기법을 접목하여 만든 부분이다.

〈그림 12〉는 새로운 사용자가 처음 방문해서 n번 장르에 대한 평가가 수행될 때까지의 과정을 나타낸 시나리오 순서도로 ⑦~⑪ 과정은 n번 반복된다.

5. 실험 및 성능평가

5.1. 실험환경 및 실험 데이터

5.1.1. EachMovie 데이터 셋

본 논문의 실험을 위한 시스템은 Java 로 작성되어 있으며 실제 실험 환경은 CPU PentiumⅣ 2.8GHz, 메모리 1.0GB 환경에서 수행되었다. 운영체제는 윈도우 XP를 사용하였고 데이터베이스는 Microsoft Access 2000을 사용하였다.

제안한 추천방법의 예측 정확도 및 연산속도를 평가하기 위해 실험 데이터로 디지털(DEC)사에서 18개월 동안 협력적 필터링 알고리즘을 연구하기 위해 사용자 선호도를 조사한 EachMovie 데이터 셋을 사용하였다.[9]

EachMovie 데이터는 72,916명의 고객들이 명시적으로 1,628의 영화에 대해 0.0에서부터 1.0까지 0.2 간격으로 평가한 2,811,983개의 선호도로 구성되어 있으며, 1996년 1월부터 1997년 9월까지 추출된 데이터이다. 영화는 총 10개의 장르(액션, 애니메이션, 외국예술, 고전, 코미디, 드라마, 가족, 공포, 스릴러)로 구성되어 있고 사용자의 인구통계학적 자료로는 성별, 연령별 정보를 포함한다. 평가정보는 0.0에서 1.0사이의 값으로 0.2의 차이를 두고 명시적으로 평가 되어 있다.

5.1.2. 실험에 사용된 정제 데이터

본 실험에서는 각 영화 정보에 포함된 장르의 조합을 가지고 76개의 서브 장르들로 분류하였다.

데이터 셋으로부터 정제 데이터를 위하여 1차적으로 무결성 검사를 수행하여 관계성이 없는 정보들을 삭제하였고 2차적으로 10세 미만과 70세 이상의 고객과 50회 이하로 평가를 수행한 고객 중심으로 데이터를 정제하여

〈표 2〉 실험데이터 정보 및 정제 데이터

테이블 명	Source 데이터	정제 데이터
Person	72,916명	9,838명
Movie	1,628개	1,628개
Voting	2,811,983건	1,113,459건
훈련집합(Training set)	-	정제 데이터의 90%
실험집합 (Test set)	-	정제 데이터의 10%

9,838명의 Person이 각 영화에 대해 평가한 1,113,459건의 선호도로 구성된 정제 데이터 집합을 구성하였다.

〈표 2〉와 같이 실험 평가를 위하여 새로운 고객을 위한 임의의 980명(10%)에 대한 Voting수를 실험집합으로 90%를 훈련 집합으로 사용하였다.

5.1.3. 실험을 위한 전제

EachMovie 데이터 셋은 고객이 한번 평가하는 시점에서 한 영화에 대해 평가한 것이 아니라 순서에 관계없이 여러 건을 평가한 데이터이며, 일정 간격을 두고 평가된 정보가 아니라 개인에 따라 서로 다른 시간 차이를 두고 평가된 실험 데이터이므로 다음과 같은 전제를 미리 세우고 실험하였다.

전제1. 훈련 집합을 1996년12월 31일로 구분하여 1단계 클러스터링을 수행하였기 때문에 새로 방문한 고객의 선호도평가 정보도 1997년 1월 1일 이후 정보들만 실험에 사용한다.

전제2. 본 논문에서는 각 고객에 대해 한번 평가된 시점을 기준으로 선호도 점수와 빈도수의 평균으로 대표 장르를 순서대로 추출하고 다시 월별로 같은 방법을 사용하여 추출한다.

전제3. 추천 사이트에 방문하는 대부분의 고객들은 개인정보 유출을 염려하여 기본정보 외에는 상세정보(취미, 선호경향, 직업 등)를 기록하지 않으려는 경향이 강하므로 가장 기본이 되는 고객 프로파일 정보만을 가지고 pre-Clustering 기법을 사용하여 1단계 클러스터링을 통해 〈표 3〉과 같이 14개의 그룹을 생성한다.

5.2 평가기준 분석 및 실험 결과

5.2.1. 실험 방법

본 실험에서는 2단계 클러스터링 방법이 성과 나이 그룹에 독립적이냐 또는 포함관계에 있느냐에 따라 기존 피어슨 상관관계를 이용한 추천과 연산시간과 예측정확도면에서 어느 정도 향상되었는지 평가한다. 그리고 장르를 10개에서 76개로 세분화하였을 경우 추천리스트를 비교함으로써 좀 더 개인화된 맞춤정보 제공을 위해서는 아이템 속성이 반영

〈표 3〉 성과 나이에 따른 그룹별 정보

그룹명	나이/성	사람수	그룹명	나이/성	사람수
그룹1	0~18(M)	709	그룹 8	10~18(F)	485
그룹2	19~24	1,449	그룹 9	19~24	885
그룹3	25~29	1,424	그룹10	25~29	901
그룹4	30~34	894	그룹11	30~34	541
그룹5	35~39	179	그룹12	35~39	194
그룹6	40~49	1,169	그룹13	40~49	344
그룹7	50 이상	520	그룹14	50 이상	144

되어야 함을 평가한다.

① 기존 GroupLens 방법 : 피어슨 상관계수식을 이용한다.

② TSC(Two-step Clustering)-독립 방법. 2단계 클러스터링이 독립적인 경우

- 새로운 고객이 들어왔을 때, 1단계에서 성과 나이에 따른 클러스터링을 수행하고 2단계에서 성과 나이의 구분 없이 독립적으로 전 시점에서 같은 장르를 본 사용자들의 데이터를 이용한다.

③ TSC(Two-step Clustering)-포함 방법. 2단계 클러스터링이 포함관계에 있는 경우

- 새로운 고객이 들어왔을 때, 1단계에서 성과 나이에 따른 클러스터링을 수행하고 2단계에서 전 시점 같은 장르를 본 고객들을 기준으로 성과 나이에 따른 장르 그룹의 데이터를 이용하였다.

5.2.2. 예측정확도 평가

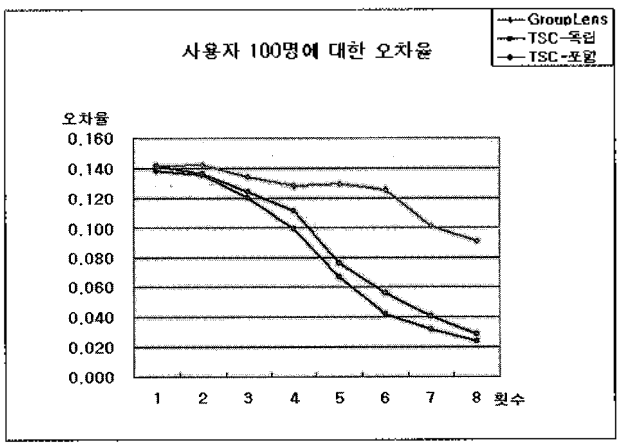
협력적 필터링을 이용한 추천시스템의 실험에서는 일반적으로 예측 정확도 평가 방법으로는 MAE(Mean Absolute Error)를 사용한다. MAE는 실험에서 발생한 평균 절대 오차 값을 말하며, 전체 예측 회수에 대해 발생한 평균 예측 오차를 의미하고 [식 4]와 같이 계산된다.

[식 2]에서 P는 사용자 상품 선호도 예측 값이며 v는 사용자 실제 평가 값이다. 예측 값과 실제 값의 차이를 구하여 그 차이를 누적하여 평균을 구함으로써 평균 얼마정도의 차이가 있는지 알 수 있는 식이다. MAE가 작을수록 추천시스템의 예측 정확도가 높음을 의미한다.

$$E = \frac{\sum |P - v|}{n}$$

[식 2] MAE(Mean Absolute Error)

본 실험에서는 10%의 테스트 집합 중 월별로 연속적인 평가 정보를 갖고 있는 사용자 24명과 한 달에서 두 달 정도의 평가 내역이



〈그림 13〉 사용자 100명에 대한 MAE 결과 그래프

빠진 100명을 추출하여 시간적 순서(월 변화)에 따라 예측 결과가 맞는지 각 사용자에 대하여 8번 반복 적용하여 평가한다.

사용자 100명에 대한 실험 결과로 〈그림 13〉에서 볼 수 있듯이 GroupLens 방법의 오차율 평균은 0.124 이고, TSC-독립 방법의 오차율 평균은 0.089, TSC-포함 방법의 오차율 평균은 0.082이다.

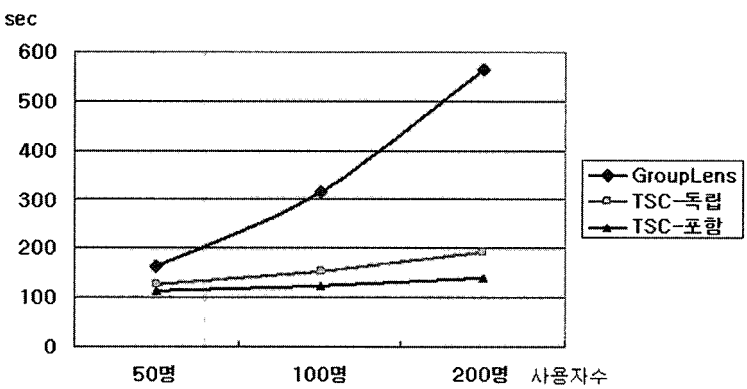
GroupLens 방법은 새로운 사용자의 경우 아이템에 평가된 자료가 거의 없으므로 약간씩 좋아지기는 하지만 예측 정확도가 고르지 못하게 나타나 초기 사용자 문제를 그대로 드러내고 있다. TSC-독립 방법과 TSC-포함 방법 모두 GroupLens 방법보다 정확도가 향상되었으나 TSC-독립방법은 오차율이 제안2보다 높다. TSC-포함 방법은 일반적은 비슷한 나이와 성(Gender)을 가진 사용자들은 취향도 비슷하므로 그 사람들을 대상으로 1단계 그룹을 형성하고 그 안에서 다시 같은 시점에 같은 장르를 좋아하는 사람들을 그룹화 하여 연산한 결과 높은 정확도를 보였다.

5.2.3. 시스템 속도 면에서의 효율성 평가

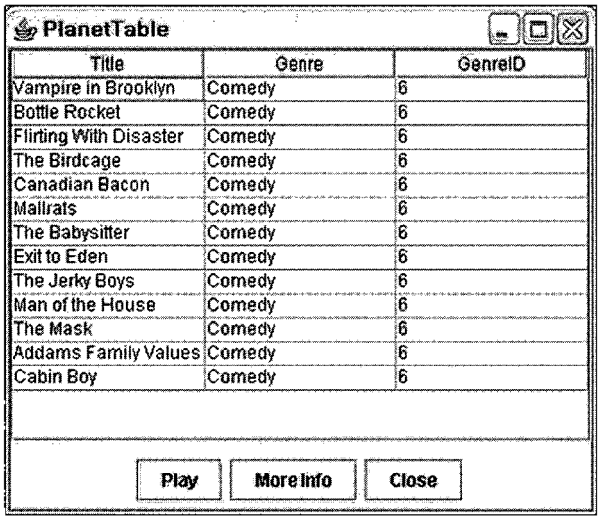
본 실험을 통해 기존 GroupLens 방법에 비해 2단계 클러스터링을 수행한 결과가 전체에 대한 계산시간보다 확연히 줄어드는 것을 알 수 있다. 〈그림 14〉에서 볼 수 있듯이 TSC-포함 방법이 TSC-독립 방법보다 공통 항목을 공유한 그룹이므로 연산시간이 많이 줄어든다. 하지만 GroupLens 방법의 경우는 사용자 수가 증가할수록 비교하는 장르수도 많아지므로 기하급수적으로 계산시간이 증가한다.

5.2.4. 추천 리스트 평가

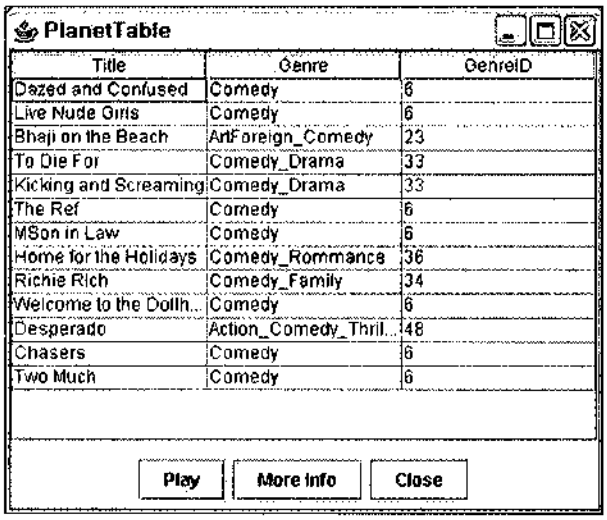
추천 리스트에 대한 평가는 각 사용자에게 적당한 아이템이 제대로 추천 되었는지를 평가하는 방법이다. 추천된 상위 n개에 대한 사용자의 평가는 사용자의 만족도로 이어지므로 추천시스템의 성능과 사이트에 대한 만족도와 신뢰도를 결정하는 중요한 요인이 된다. 추천 리스트에 대한 평가 방법으로는 〈그림 15〉 과 같이 76개 장르로 구분했을 때와 〈그림 16〉과 같이 10개 장르를 구분했을 때 리스



〈그림 14〉 사용자수에 따른 계산 시간



〈그림 15〉 76개 장르 추천 리스트



〈그림 16〉 10개 장르 추천 리스트

트 후보들을 비교한다.

첫 번째는 코미디 장르만으로 영화가 구성되었으나, 두 번째 화면은 코미디 장르와 조합된 모든 영화 리스트가 보여 진다. 따라서 좀 더 개인화된 맞춤정보 제공을 위해서는 아이템 항목을 세분화하여 그 속성이 반영되어야 함을 알 수 있다.

6. 결론 및 향후 연구과제

전자상거래 환경에서 점차 방대해지는 사용자와 상품에 관한 정보를 바탕으로 하는 추천시스템이 가지는 문제점은 이미 많은 지적을 받아 왔다. 사용자 수가 증가함에 따라 추천 항목을 결정하는데 걸리는 시간이 증가하는 확장성(scalability) 문제와 새로운 사용자의 경우와 같이 곡개에 대한 선호도 정보가 부족할 경우 추천 정확도가 저하되는 희박성(saparsity) 문제가 그것이며, 이들을 해결하기 위한 많은 연구와 실험이 이어졌으나 아직도 개선의 여지가 남아 있는 상황이다.

본 논문에서는 추천시스템에서 가장 보편적으로 보이고 있는 협력적 필터링(collaborative filtering) 방법을 사용하여 이러한 문제들을 해결하기 위한 방안을 제시하였다. 대부분의 협력적 필터링 시스템들은 사용자간의 유사도를 구하는데 있어서 코사인함수나 피어슨 상관계수식을 이용하기 때문에 아이템수가 많아질수록 사용자가 관련된 정보를 얻는데 어느 정도 한계가 있다. 따라서 두 사용자간에 선호도를 표시할 확률은 적어지게 되고, 상관관계를 비교할 아이템 수는 증가하게 된다.

본 논문에서는 협력적 필터링 방법을 사용하면서 시간에 따른 성향 변화를 고려하고 2단계에 걸쳐 클러스터링 하는 기법을 사용하여 희박성 문제와 확장성 문제를 해결하고 예측정확도를 개선하였다.

본 논문에서 아직 해결되지 않은 부분으로 앞으로 수행되어야 할 연구과제는 다음과 같다.

- 사용자의 선호도를 보다 잘 표현할 수 있는 데이터를 선택 연구하여 보다 높은 성능을 보장할 수 있는 시스템을 구축하는 것이 필요하다.
- 본질적으로 협력적 추천시스템이 안고 있는 결여 데이터 문제를 해결 혹은 완화할 수 있는 방안에 대한 연구가 필요하다.
- 대상고객과 유사한 히스토리를 가지는 특정 N명의 이웃이 가진 정보를 바탕으로 추천이 이루어지므로 국소적 추천에 머물게 되고, 나머지 이웃들에게서 이끌어낼 수 있는 전역적 추천을 놓칠 수 있으므로 그에 대한 방안에 대한 연구가 필요하다.
- 시간 의존적 성향을 반영하기 위한 Dynamic Learning 기법에 대한 확대 연구가 필요하다.

참 고 문 헌

[1] kai Yu. Anton Schwaighofer, Volker Tresp, Xiaowei Xu and Hans-peter Kriegel. "Probabilistic Memory-Based Collaborative Filtering", IEEE Transactions on knowledge and data Engineering, vol. no.1 January 2004.

[2] Lyle H. Ungar and Dean P. Foster, "Clustering Methods for Collaborative Filtering", In Proceedings of the AAAI-98 Workshop on Recommender Systems, 1998

[3] 정경용, 최성용, 임기욱, 이정현, "베이지안 추정치가 부여된 유사도 가중치와 연관 사용자 군집을 이용한 선호도 예측시스템", 정보과학회 논문지 제30권 제4호, 2003년 4월.

[4] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, John T. Redle, "Application of Dimension ality Reductionn Recommender Sy stem - A Case Study", ACM WebKDD 200 Web Mining for E-Commerce Workshop, 2000.

[5] Lyle H. Ungar and Dean P. Foster, "Clustering Methods for Collaborative Filtering", Proceeding of the 1998 Workshop on Recommendation Systems, pp.114-129, 1998.

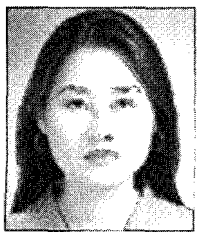
[6] Resnick, P., Iacovou, N., Such ak, M., Berg strom, P., and Riedl, J., "GroupLens : An Open Architecture for Collaborative Filt ering of Netnews", Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, 1994.

[7] Herlocker J., Konstan J., Borchers A., and Riedl J., "An algorithmic framework for performing collaborative filtering", Proceedings of the 22nd Conference on Research and Development in Information Retrieval, 1999.

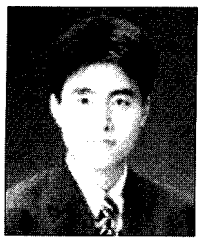
[8] Breese J., Heckerman D. and Kadie C., "Empirical analysis of predictive algorithms for collaborative filtering", Proceedings of the 14th Conference of Uncertainty in Artificial Intelligence, 1998.

[9] http://www.cs.umn.edu/Research/GroupLens/

저 자 소 개



부종수 (E-mail : bujongsu@cnu.ac.kr)
1992. 2 울산대학교 전산학과 (이학사)
2005. 2 충남대학교 컴퓨터과학과 (이학석사)
현재 충남대학교 누리사업단 조교
관심 분야 개인화, 추천시스템, e-비지니스



홍종규 (E-mail : jkhong@cs.cnu.ac.kr)
2004. 2 충남대학교 컴퓨터과학과 (이학사)
현재 충남대학교 컴퓨터공학과 석사과정
관심 분야 지능형 추천 알고리즘, e-비지니스

박원익 (E-mail : wonik78@cs.cnu.ac.kr)
2004. 2 충남대학교 컴퓨터과학과 (이학사)
현재 충남대학교 컴퓨터공학과 석사과정
관심 분야 Mobile Data Management, 개인화, 추천시스템

김 룡 (E-mail : ryong@cs.cnu.ac.kr)
2000. 2 충남대학교 컴퓨터과학과 (이학사)
2003. 2 충남대학교 컴퓨터과학과 (이학석사수료)
현재 주식회사 코이스트 대표이사 / 대전보건대학겸임교수
관심 분야 무선 인터넷, B2C, B2B, 정보 검색

김영국 (E-mail : ykim@cnu.ac.kr)
1985. 2 서울대학교 계산통계학과 (학사)
1987. 2 서울대학교 계산통계학과 (이학석사)
1995. 5 버지니아대학교 컴퓨터과학과 (공학박사)
1995. 3 ~ 1995. 8 핀란드 VTT 연구소 - 방문연구원
1995. 9 ~ 1996. 2 노르웨이 SINTEF DELAB - 방문연구원
1996. 3 ~ 현재 충남대학교 전기정보통신공학부 부교수
2002. 8 ~ 2003. 7 UC Davis - 방문교수
관심 분야 실시간 데이터베이스, 멀티미디어 및 모바일 정보시스템, 전자상거래 시스템