일반논문-09-14-5-08

순차 패턴 마이닝 기법을 이용한 개인 맞춤형 TV 프로그램 스케줄러

표 신 지[a], 김 은 희[b], 김 문 철[b]‡

A Personalized Automatic TV Program Scheduler using Sequential Pattern Mining

Shinjee Pyo[a], Eunhui Kim[b], and Munchurl Kim[b]‡

요 약

방송 프로그램 콘텐츠들의 증가와 콘텐츠 접근 방법의 다양화로 따라 사용자는 기존의 단순한 방송 시청 환경에서 보다 복합적인 환경에서 다양한 콘텐츠를 접할 수 있게 되었다. 따라서 사용자는 익숙지 않은 다양한 콘텐츠들 중에서 자신이 시청하기 원하는 콘텐츠를 찾고 그것들을 원하는 시간에 시청하기 위해 전보다 많은 노력을 기울이게 되었다. 또한 사용자는 대체로 자신만의 일관성 있는 시청 패턴으로 프로그램을 시청한다. 본 논문에서는 사용자의 개인적인 시청 특성을 발견하여 사용자의 수고를 줄이고 프로그램 시청의 편의성을 제공하기 위해 순차 패턴 마이닝 기법을 이용하여 개인 맞춤형 TV 프로그램 스케줄러를 제안한다. 이를 위해 개인 맞춤형 TV 프로그램 스케줄 추천 시스템을 제안하였으며, 사용자들의 TV 프로그램 시청 기록을 바탕으로 TV시청 환경에 적합한 순차 패턴 마이닝 기법을 제안하였다. 또한 개인 사용자의 암시적인 선호도를 추출하여 TV 프로그램 추천에 적용, 개인 맞춤형 TV 프로그램 스케줄을 구성하여 추천할 수 있도록 하였다. 이러한 TV 프로그램 스케줄 추천 시스템은 향후 IPTV의 VoD 특성을 고려한 프로그램 스케줄 추천 시스템으로 확장가능하다.

Abstract

With advent of TV environment and increasing of variety of program contents, users are able to experience more various and complex environment for watching TV contents. According to the change of content watching environment, users have to make more efforts to choose his/her interested TV program contents or TV channels than before. Also, the users usually watch the TV program contents with their own regular way. So, in this paper, we suggests personalized TV program schedule recommendation system based on the analyzing users' TV watching history data. And we extract the users' watched program patterns using the sequential pattern mining method. Also, we proposed a new sequential pattern mining which is suitable for TV watching environment and verify our proposed method have better performance than existing sequential pattern mining method in our application area. In the future, we will consider a VoD characteristic for extending to IPTV program schedule recommendation system.

Keyword : Recommendation system, Scheduler, Sequential pattern mining

b) 한국과학기술원 정보통신공학과
Dept. of Information and Communications Engineering, Korea Advanced Institute of Science and Technology
b) 한국과학기술원 전기및전자공학과
Dept. of Electrical Engineering, Korea Advanced Institute of Science and Technology
‡ 교신저자 : 김문철(mkim@ee.kaist.ac.kr)
* 본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음[A1100-0801-3015, Development of Open-IPTV Technologies for Wired and Wireless Networks].
· 접수일(2009년7월1일),수정일(2009년9월4일),게재확정일(2009년9월21일)

Ⅰ. 서 론

방송과 통신의 융합으로 인해 수많은 콘텐츠들이 사용자들에게 제공 가능해졌고 디지털 기기의 발달과 영상 편집 기술의 대중화로 UCC와 같은 다양한 콘텐츠가 생성되었다. 기존의 TV 시청환경은 사용자가 자신이 원하는 프로그램을 채널 검색 등을 통하여 선택하고 시청하는 형태였다. 하지만 채널 및 콘텐츠의 다양화와 인터넷의 발달로 사용자는 매우 다양한 콘텐츠를 접할 수 있게 되었다. 또한 사용자는 대체로 자신의 일정한 순서를 가지고 프로그램을 시청한다. 이러한 시청 환경에서 사용자는 이전보다 다양해진 프로그램 중에서 자신이 선호하고 자신에게 친숙한 콘텐츠를 찾는 것에 대해 부담을 느낄 수 있으며 연속성 있는 프로그램 시청에 어려움을 느낄 수 있다. 따라서 사용자에게 적합한 콘텐츠와 시청 스케줄을 추천해 줄 수 있는 추천 시스템이 요구된다.

이러한 추천 시스템은 사용자의 TV 시청 기록을 기반으로 시청 패턴을 분석하고 사용자 선호도 정보를 추출하는 과정을 통해 이루어 질 수 있다. 시청 패턴 분석을 위해 데이터 마이닝의 한 분야인 순차 패턴 마이닝 을 적용하여 사용자의 의미 있는 패턴을 추출하고 개인 사용자의 선호도를 고려하여 개인 맞춤형 프로그램 스케줄을 추천한다.

본 논문은 다음과 같이 구성되어 있다.

Ⅱ장은 기존의 순차 패턴 마이닝 기법을 이용한 추천 시스템에 대한 내용과 순차 패턴 마이닝 기법에 대한 연구에 대해 정리한다. Ⅲ장은 본 논문에서 제안하는 개인 맞춤형 프로그램 스케줄러 추천 시스템의 구조와 각각의 메커니즘에 대해 설명한다. Ⅳ장에서는 본 논문에서 제안한 추천 방법에 대해 실험한 결과에 대해 설명하고 Ⅴ장에서는 본 논문의 결론에 대해 설명한다.

Ⅱ. 관련 연구

순차 패턴 마이닝(Sequential Pattern Mining)기법은 사용자의 특정한 행동에 대해서 순차적인 패턴을 찾아내는 기법이다[1]. 이 기법은 데이터 마이닝의 한 방법으로서 시간적인 순서를 고려하여 연관 규칙을 찾아내는 기법이다. 이러한 기법을 이용하여 사용자의 구매 패턴을 발견한다거나 웹 사이트 방문 패턴을 발견하여 특정 물품구매 또는 웹 사이트 방문 이후의 행동을 예측하여 사용자에게 물품이나 웹 사이트를 추천할 수 있다. 기존의 연구에서는 이러한 순차 패턴 마이닝 기법을 인터넷 쇼핑몰에서의 사용자의 구매 데이터베이스에 적용하여 사용자의 순차적인 구매 패턴을 추출하고 이를 통해 상품을 추천하는 접근 방법을 사용하였다[3],[4].

웹 로그 분석을 이용한 추천 에이전트 개발에 관한 논문[4]에서는 전자상거래에서의 사용자의 물건 구매 데이터베이스를 기반으로 사용자가 특정 물품 구매 이후에 어떤 물품을 구매할 것인지 예측하여 추천하는 추천 에이전트를 구현하였다. 예측하는 방법으로는 사용자들의 구매 데이터베이스에서 연관 규칙(Association rule)을 이용하여 빈발적으로 발생하는 구매 패턴을 발견하고, 순차 패턴 마이닝 기법을 이용하여 시간적 순서를 갖는 구매 순서 패턴을 추출한다. 이러한 데이터베이스의 분석을 통한 패턴 추출에 대한 부분은 분석 에이전트에서 수행한다. 추천 에이전트에서는 분석 에이전트에서 추출한 연관 규칙들과 순차 패턴들을 이용하여 실제 사용자에게 물품을 추천하는 역할을 담당한다. 추천 에이전트의 추천 형태는 사용자에게 여러 개의 추천 리스트를 제공하여 사용자가 선택하도록 하는 형태이다. 실험을 위해서 이 논문에서는 200명의 회원을 보유하고 있는 실험용 인터넷 쇼핑몰을 구축하고 이 200명의 회원으로 하여금 각각 5건씩의 거래를 하도록 하였다. 실험 데이터에서 연관 규칙과 순차 패턴을 발견하였으며, 순차 패턴을 이용하였을 때 보다 유용한 규칙들을 많이 발견할 수 있었다.

사용자들의 시간적 순서에 따른 특정 패턴을 발견하기 위한 순차 패턴 마이닝 기법에는 AprioriAll[1], GSP[5], PrefixSpan[2], FreeSpan[6]등 다양한 종류의 기법이 연구되어 왔다. 이 중 PrefixSpan은 계산량이 상대적으로 적어 빠른 계산 속도를 갖는다. PrefixSpan 방법은 다음과 같은 여섯 단계의 과정을 거쳐 순차 패턴을 생성한다. 먼저, 사용자들의 구매 내역을 시간적 순서에 따라, 사용자 별로 시퀀스로 정렬하는 과정을 거친다. 이를 통해 사용자는 일정 기간 동안의 구매 아이템들을 시퀀스로 갖게 된다. 두 번째 단계는

첫 번째 단계에서 생성된 시퀀스들을 기반으로 그 시퀀스에 자주 등장하는 아이템을 선별하는 단계이다. 선별하는 방법은 최소 지지도(minimum support)를 이용하여 아이템의 구매 횟수가 임의로 정한 최소 지지도를 넘는 아이템을 빈발 아이템으로 선별한다. 즉, 시퀀스를 구성하는 총 사용자가 가 100명일 때, 100명 중 아이템 A를 구매한 사람의 수가 50명이라고 하고, 임의로 정한 최소 지지도가 전체 사용자의 25%라고 하였다면, 아이템 A는 100명의 25%인 25명을 넘었기 때문에 빈발 아이템이 된다. 두 번째 단계에서 이와 같이 빈발 아이템들을 선별한 후에는 이 빈발 아이템들을 포함하고 있는 시퀀스들을 골라내는 단계를 거친다. 두 번째 단계에서 선별된 빈발 아이템들이 A, B, C라고 한다면, 이 단계에서는 A, B, C를 포함하고 있는 시퀀스만 골라내는 단계이다. 네 번째 단계는 세 번째 단계에서 골라낸 시퀀스들을 각 빈발 아이템 A, B, C를 포함하는 시퀀스들로 나누어 투사(projection)과정을 거쳐 A의 투사 데이터베이스(projected database), B의 투사 데이터베이스(projected database), C의 투사 데이터베이스(projected database)를 생성하는 단계이다. 여기서 투사 데이터베이스(projected database)는 각 빈발 아이템 A, B, C로 투사한 시퀀스들의 집합이라는 의미를 갖으며, 투사는 특정 아이템(빈발 아이템 A, B 또는 C)으로 시퀀스를 잘라내어 투사하는 것을 뜻한다. 즉, <E-G-A-D-F>라는 시퀀스가 주어졌을 때, 이 시퀀스는 A를 포함하고 있으므로, A에 대해 투사를 수행하면, <D-F>로 투사 시퀀스가 구성된다. 이러한 아이템 A에 대한 투사 시퀀스들의 집합이 A의 투사 데이터베이스가 된다. 마지막 단계는 이 투사 데이터베이스에서 다시 빈발적으로 나타나는 아이템을 골라내는 것이다. 즉 A의 투사 데이터베이스에 존재하는 A 이후의 다양한 시퀀스들에서 자주 나타나는 아이템을 앞서 언급한 최소 지지도를 이용하여 선별해낸다. 예를 들어, A의 투사 데이터베이스에서 최소 지지도를 만족하는 아이템이 D와 F였다고 하면, <A-D>, <A-F> 와 같은 길이-2의 순차 패턴이 생성된다. 길이가 더 긴 패턴을 생성하기 위해서는 생성된 길이-2의 시퀀스를 이용하여 투사를 수행하여 길이-2의 순차 패턴에 대한 투사 데이터베이스를 생성하고 다시 빈발 아이템을 선별하는 과정을 거쳐 길이-3, 길이-4, 그 이상인 순차적인 패턴을 생성할 수 있다.

Ⅲ. 개인 맞춤형 프로그램 스케줄러 추천 시스템

1. 시스템 구조

본 논문에서는 사용자들의 누적된 시청 기록을 이용하여 순차적인 패턴을 발견하고 또한 사용자의 암시적인 선호도를 추출하여 사용자에게 적합한 프로그램 스케줄을 제공하는 추천 시스템을 제안하였다. 그림 1 은 본 논문에서 제안한 추천 시스템을 나타낸다.

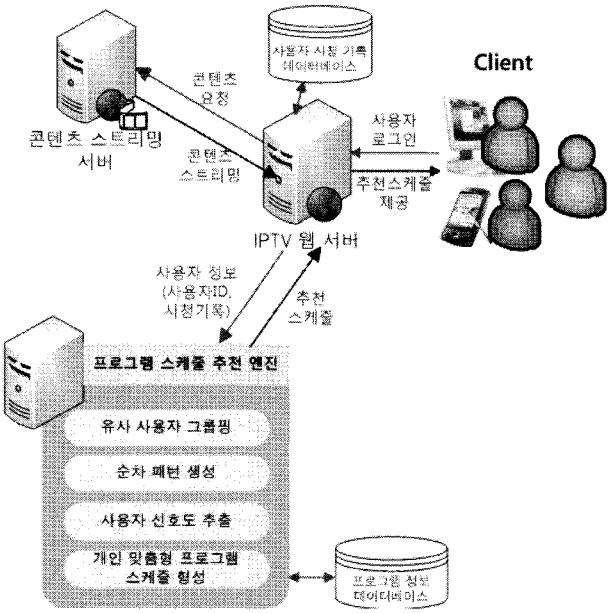

그림 1. 개인 맞춤형 프로그램 스케줄 추천 시스템 구조
Fig. 1. A recommendation system for personalized TV program scheduler

사용자는 TV나 PC 혹은 PDA등의 단말기를 통해 IPTV 서비스에 로그인 한다. 서비스를 관리하는 IPTV 웹 서버는 이 로그인 정보를 추천 엔진에 전달하고, 추천 엔진은 로그인 사용자를 위한 추천 스케줄을 추출한다. 사용자가 추천 메뉴에 대한 서비스를 요청하면 웹 서버는 추천 엔진에게 사용자의 요청을 전달하고 추천 엔진은 그 사용자의 추천 프로그램 스케줄을 웹 서버에 전달, 웹 서버에서 추천 프로그램 스케줄 정보를 이용하여 서비스를 요청한 사용자에

맞게 웹 페이지를 구성하여 프로그램 스케줄을 표시하여 제공한다. 추천 엔진은 사용자의 시청 기록 데이터베이스와 연결되어 있으며 시청 기록 데이터베이스는 IPTV 웹 서버와 연결되어 있다. 따라서 사용자가 프로그램을 시청하기 시작할 때와 그 프로그램을 시청 종료 할 때, 웹 서버는 시작 시점과 종료 시점(날짜/시간/요일)을 사용자 시청 기록 데이터베이스에 기록한다. 또한 사용자가 시청한 프로그램의 제목, 장르, 방영 채널, 방영 시간, 출연 배우에 대한 정보도 함께 기록하여 사용자가 어떤 프로그램을 얼마만큼 시청하였는지에 대해 기록한다. 만일 시청한 프로그램이 방송 콘텐츠가 아닌 VoD 콘텐츠라고 하면 방영 시간이 따로 정해져 있지 않고 사용자가 원하는 시간에 언제든지 시청 가능하다. 따라서 VoD나 UCC 등과 같은 콘텐츠의 경우 프로그램 방영 시간에 대한 정보는 기입하지 않는다. 사용자의 시청 기록 데이터베이스에는 사용자의 id 별로 사용자가 시청한 프로그램들에 대해서 각각의 시청 날짜, 시청한 시간, 프로그램을 방영한 채널, 프로그램의 장르, 프로그램의 방영 시작/종료 시간 등이 기록되어 저장되어 있다.

이러한 사용자의 시청 행동에 대해 저장하는 과정을 거쳐 수집된 사용자의 누적 시청 기록을 이용하여 오프라인으로 사용자의 프로그램 스케줄러를 생성한다. 먼저 누적된 시청 기록을 이용하여 사용자의 암시적인 선호도를 추론할 수 있다. 이러한 사용자 개인의 선호도 정보를 바탕으로 유사한 사용자들로 그룹핑을 수행한다. 생성된 유사 시청 사용자 그룹 내에서 순차적인 시청 프로그램 패턴을 추출한다. 또한 사용자 그룹 내의 사용자들의 다양한 선호도 정보를 사용하여 각 사용자 별 시청 프로그램 패턴 추천 순위를 결정하여 개인 프로그램 스케줄러를 생성한다. 오프라인으로 생성된 스케줄러는 사용자가 로그인 했을 때의 시간정보에 따라 온라인으로 시청 가능한 프로그램 스케줄을 선별하여 사용자에게 추천해준다.

추천 엔진의 각 단계에 대해서는 다음 절에서 자세히 설명한다.

2. 유사 사용자 그룹핑

순차 패턴 마이닝 기법은 사용자 그룹 내에서 사용자들의 누적된 구매 기록 또는 검색 기록 등을 이용하여 사용자 그룹의 순차적인 패턴을 찾아내는 기법이다. 따라서 사용자들의 순차적인 시청 프로그램 패턴을 찾아내기 위해 사용자 그룹으로 범위를 지정하여 그 그룹 내에서의 빈발적인 시청 패턴을 찾아내는 접근 방법을 사용한다. 사용자 그룹으로 나눌 때에 임의로 사용자 그룹을 지정하는 것 보다 시청하는 프로그램이 유사한 사용자들로 사용자 그룹을 형성하는 것이 보다 효과적인 시청 패턴 발견방법이 될 수 있다. 본 절에서는 유사한 프로그램을 시청하는 사용자들로 구성된 사용자 그룹을 형성하는 방법에 대해 논의한다.

다양한 사용자들이 있을 때, 어떤 사용자들이 유사한 시청 습관을 갖는가를 판단하기 위해서 사용자의 개인별 시청 프로그램에 대한 선호도를 추론한다. 선호도는 사용자가 시청한 각 프로그램의 시청 시간 정보를 기반으로 계산할 수 있다. 개인의 프로그램 별 선호도를 위한 계산식은 다음과 같다.

$$PP_{i,u_a} = \frac{1}{M}\sum_{k=1}^{M}\frac{wt_{i,u_a,k}}{d_{i,k}} \quad (1)$$

여기서 PP_{i,u_a}는 사용자 u_a의 TV 프로그램 i에 대한 시청 정도를 나타낸 값으로, 프로그램 i가 방영할 때마다의 방영 시간 대비 시청 시간을 모두 더해 총 방영 횟수로 나눈 값이다. M은 프로그램 i의 총 방영 횟수를 나타내고, $d_{i,k}$는 프로그램 i의 k 번째 방영 날짜의 방영 시간을 나타낸다. $wt_{i,u_a,k}$ 는 사용자 u_a의 프로그램 i에 대한 시청 시간을 나타낸다. 따라서 시청 시간을 방영 시간으로 나눈 값을 더하여 사용자의 특정 프로그램 i에 대한 선호도를 암시적으로 계산할 수 있다.

이 수식을 이용하여 모든 사용자에 대해 사용자가 시청한 프로그램 각각의 선호도 값을 계산한다. 본 논문에서는 기준이 되는 사용자 1명을 임의로 선정하여 이 사용자의 PP_{i,u_a}가 큰 순서대로 10개의 프로그램을 선별하고, 큰 순서대로 프로그램의 PP_{i,u_a}값을 벡터를 구성하는 요소로 정하여 가장 큰 프로그램의 PP_{i,u_a}값을 벡터를 구성하는 첫 번째 요소로, 10번째로 큰 PP_{i,u_a}값을 벡터를 구성하는 열

번째 요소로 다음과 같이 벡터를 생성하였다.

$$v_{u_a} = \left[w_{1,u_a}, w_{2,u_a}, \cdots, w_{10,u_a}\right]^T, w_{1,u_a} = MAX\left\{PP_{i,u_a}\right\} \quad (2)$$

이렇게 기준 사용자의 시청 프로그램에 대한 벡터를 형성한 후에, 기준 사용자가 가장 좋아한 10개의 프로그램들, 즉 기준 벡터를 구성하는 프로그램들에 대한 다른 사용자들의 프로그램 선호도 값을 가지고 다른 사용자들의 프로그램 선호도 벡터로 정하였다. 따라서 기준 벡터를 구성하는 프로그램에 대한 사용자의 선호도의 유사성을 측정하여 유사 사용자 그룹을 형성하였다. 만일 어떤 사용자가 기준 벡터를 구성하고 있는 특정 프로그램에 대해 시청하지 않았다면, 그 프로그램에 대한 선호도 값은 0으로 기록한다. 즉 유사 사용자 그룹은 임의로 정한 한 사용자와 비슷한 프로그램 시청 습관을 갖는 사용자들로 구성된 그룹이다. 기준 벡터와 다른 사용자들의 벡터의 유사성을 계산하기 위해서 다음 수식과 같은 벡터 공간 모델(vector space model)을 이용한다.

$$\cos\theta = \frac{v_{u_a} v_{u_b}}{\| v_{u_a} \| \| v_{u_b} \|}, v_{u_a} = \left[w_{1,u_a}, w_{2,u_a}, \cdots, w_{10,u_a}\right]^T \quad (3)$$

이와 같이 두 벡터간의 유사성을 $\cos\theta$값으로 구하여 기준 벡터와 가장 유사한 30명의 사용자들로 유사 시청 사용자 그룹을 형성하였다.

3. 순차 패턴 형성 단계

2절에서 생성한 유사 시청 사용자 그룹 내의 순차 패턴을 형성하는 단계이다. 이 단계에서는 TV 프로그램 시청 환경에 적합한 시퀀스 표현 방법을 제안하고 이에 따른 순차 패턴 마이닝 기법을 제안한다.

3.1 사용자의 프로그램 시청 시퀀스의 표현 방법

기존의 순차 패턴 마이닝 기법에서 주로 적용한 아이템 구매 환경에서는 아이템의 구매 횟수를 이용한 빈발 아이템 선별 방법이 적합하였다. 여러 사용자들에게 구매되는 아이템을 의미있는 아이템으로 추출하여 그 아이템 이후에 구매된 아이템들을 관찰함으로써 순차적인 패턴을 발견할 수 있었다. 하지만 TV프로그램 시청 환경의 경우에는 사용자의 프로그램 시청 시퀀스를 단순히 $<A-B-C>$ (A,B,C는 시청 프로그램)로 하기엔 방송 환경에서 포함해야 할 의미 있는 정보들을 포함할 수 없다. 프로그램 시청자는 물품 구매와 같이 물품인 프로그램을 선택하는 것뿐만 아니라 선택한 후에 일정시간 동안 프로그램을 시청한다. 또한 본 논문에서 사용한 지상파 프로그램은 방영 시간이 정해져 있고, 방영 시간이 지나면 시청할 수 없다. 따라서 본 논문에서는 이러한 사용자의 프로그램 시청 시간과 프로그램 방영 시간에 대한 정보를 고려하여 사용자의 시청 프로그램 시퀀스를 구성하고, 형성된 시퀀스로 빈발 프로그램을 추출하는 방법을 제안하였다. 다음 식은 사용자의 프로그램 시청 시간 정보와 프로그램 방영 시간 정보를 고려한 시퀀스이다.

$$User\ u_a{}'s\ sequence: < p_{i,wq_{i,u_a},bt_i} - p_{j,wq_{j,u_a},bt_j} > \quad (4)$$

위의 식은 두 가지 프로그램, p_i, p_j을 시청한 사용자 u_a의 시청 시퀀스 이다. 인덱스 i, j는 프로그램 타이틀에 대한 인덱스이고, wq_{i,u_a}는 사용자 u_a의 프로그램 p_i의 방영시간 대비 시청 시간으로 다음과 같은 수식으로 표현된다.

$$wq_{i,u_a} = \frac{wt_{i,u_a}}{d_i} \quad (5)$$

wt_{i,u_a}는 사용자 u_a의 프로그램 p_i의 시청시간을 분 단위로 나타낸 값이고, d_i는 프로그램 p_i의 프로그램 방영시간이다. 예를 들어 방영 시간이 1시간인 프로그램 p_m을 사용자 u_a가 30분 동안 시청하였다고 하면, 이에 대한 wq 값은 다음과 같이 나타낼 수 있다.

$$wq_{m,u_a} = \frac{wt_{m,u_a}}{d_m} = \frac{30}{60} = 0.5 \quad (6)$$

따라서 wq값을 이용하여 특정 프로그램에 대한 사용자

의 시청 정도를 계산하여 시청 시퀀스를 표현하는데 함께 사용할 수 있다. 다음 항목은 프로그램 p_i의 방영 시작시간을 나타내는 bt_i이다. bt_i는 프로그램의 시작 시간을 다음과 같은 형태로 나타낸다.

$$bt_i = HourOfBST(p_i) + \frac{MinuteOfBST(p_i)}{60} \quad (7)$$

$HourOfBST(p_i)$는 프로그램 p_i의 방영 시작 시간에서 Hour에 해당하는 값을 나타내고, $MinuteOfBST(p_i)$는 Minute에 해당하는 값을 나타내어, bt_i는 프로그램의 방영 시작 시간을 하나의 실수 값으로 표현하도록 하였다. 즉 18시 25분에 시작하는 프로그램p_i의 bt_i는 다음과 같이 나타낼 수 있다.

$$bt_i = 18 + \frac{25}{60} \approx 18 + 0.4 = 18.4 \quad (8)$$

이러한 방송 시청 환경을 고려한 시청 시간 정보, 프로그램 방영 시작 시간 정보를 이용한 시퀀스는 사용자들의 시청 기록 데이터베이스에서 각 사용자 별로 나뉜 후 각 날짜별로 나뉘어 각 사용자의 날짜 별 시퀀스로 표현되며 이를 이용하여 보다 적합한 프로그램 시청 순차 패턴을 추출할 수 있다.

다음 절에서는 제안한 시퀀스를 이용한 순차 패턴 마이닝 기법에 대해 설명한다.

3.3.2 순차 패턴 마이닝

본 논문에서 기본적으로 적용한 방법은 순차 패턴 마이닝 기법 중 하나인 PrefixSpan방법이다. PrefixSpan방법은 데이터베이스에서 빈발적으로 나타나는 아이템들을 선별하고, 그 아이템들을 prefix로 사용하여 순차적인 패턴을 발견하는 방법이다. 빈발 적으로 나타나는 아이템을 선별하는 방법은 데이터베이스에 존재하는 아이템들의 빈도, 즉 사용자들의 아이템 구매 데이터베이스의 경우 아이템 각각의 구매 횟수를 이용하여 선별한다. 선별하는 기준은 최소 지지도(minimum support)값을 이용하여 전체 사용자 수의 25%(혹은 20%, 임의로 정하는 값) 이상의 구매 횟수를 갖는 아이템을 빈발 아이템으로 선별한다. 이러한 PrefixSpan 기법 그대로를 사용자의 프로그램 시청 환경에서 적용하기에는 사용자의 프로그램 시청 시간 정보와 같은 방송 시청 환경 특성을 고려하지 못해 보다 의미있고 정확한 패턴을 추출할 수 없다. 따라서 앞 절에서 정의한 사용자의 프로그램 시청 시퀀스를 이용하여 시청 프로그램 패턴을 추출하는 방법에 대해 알아보도록 하겠다.

앞서 정의한 시퀀스에는 기존의 PrefixSpan에서는 고려하지 않은 두 가지 정보가 추가적으로 포함되어 있다. 이러한 정보들을 이용하여 사용자들이 충성도를 가지고 빈발적으로 시청한 프로그램을 prefix 프로그램으로 선별할 수 있다.다음 수식은 각 요일의 prefix 프로그램을 선별하기 위한 각 프로그램 별 중요도 값을 나타낸다.

$$Q_i = \frac{1}{F_i} \sum_{n=1}^{N} wq_{i,n} \quad (9)$$

F_i는 일정 기간 동안의 프로그램 p_i의 방영 횟수를 뜻하고, $\sum_{n=1}^{N} wq_{i,n}$는 프로그램 p_i에 대한 $wq_{i,n}$를 그룹 내의 전체 사용자 N명에 대해서 모두 더한 값을 나타낸다. 즉, Q_i 값은 프로그램 p_i에 대한 전체 사용자의 충성도를 나타낸다. 만일 모든 사용자가 프로그램 p_i가 방영할 때마다 처음부터 끝까지 모두 시청 하였다면, Q_i는 전체 사용자의 수인 N이 되고, 반대로 모든 사용자가 한 번도 시청하지 않았다면 Q_i는 0이 될 것이다. 대부분의 지상파 방송 프로그램들은 일주일을 단위로 반복되므로 사용자들의 시청 프로그램을 요일별로 나누어 각 요일별 프로그램들의 Q_i값과 시청 횟수를 이용하여 요일별 빈발 시청 프로그램을 선별한다. 선별을 위한 문턱(threshold)값으로는 Q_i값이 특정 값(예 2) 이상이고 시청 횟수가 요일의 전체 시퀀스의 특정 %(예 20%) 이상인 프로그램을 빈발 시청 프로그램으로 선별한다. 이러한 문턱값으로 요일 별 빈발 프로그램을 결정하여 패턴 형성을 위한 prefix들을 생성하고. 이를 이용하여 투사를 하게 된다. 투사는 prefix로 시퀀스를 나누는 것을 말하는 데 시퀀스를 투사한 결과를 postfix라고 하며 이러한 과정을 다음과 같이 표현할 수 있다.

$$prefix : B$$
$$Seq1 : < A-B-C-D-E >$$
$$Seq2 : < A-B-D-F-G >$$
$$Seq3 : < B-D-E >$$
$$\Downarrow projection\ with\ prefix\ B \quad (10)$$
$$postfix1 : < C-D-E >$$
$$postfix2 : < D-F-G >$$
$$postfix3 : < D-E >$$

즉, prefix B를 포함하는 시퀀스를 골라내서 prefix B이후의 시퀀스들만 남겨 postfix로 삼는다. 이렇게 생성된 postfix들의 집합을 prefix B의 투사 데이터베이스라고 하며, 각 요일별 빈발 시청 프로그램들 각각에 대해 모두 실행하여 각 프로그램에 대한 투사 데이터베이스들을 생성한다. 이 투사 데이터베이스내 에서 앞서 설명한 빈발 프로그램을 선별하는 과정을 동일하게 실행한다. 그 결과, 요일별 빈발 프로그램별 투사 데이터베이스내의 빈발 시청 프로그램들이 선별되어, 길이-2의 시청 프로그램 패턴이 발견된다. 즉, 월요일의 빈발 시청 프로그램으로 선별된 프로그램이 A라고 하면, A에 대해 투사를 수행한 결과로 A의 투사 데이터베이스가 생성된다. 이 A의 투사 데이터베이스내에서 문턱값을 만족시키는 프로그램들로 B, C, D라는 프로그램들이 빈발 프로그램으로 선별되었다고 하면, 월요일의 순차 시청 프로그램 패턴으로 A-B, A-C, A-D 세 종류의 길이-2의 패턴이 발견된다.

실제로 본 논문에서 사용한 실험 데이터를 기반으로 유사 시청 사용자 그룹에 대해 각 요일별 순차 시청 프로그램 패턴을 추출한 결과는 표 1과 같다.

요일별 빈발 프로그램의 투사 데이터베이스내의 빈발 프로그램을 prefix로 하여 투사과정을 다시 한 번 수행하게 되면 해당하는 prefix의 투사 데이터베이스가 생성되고, 투사 데이터베이스 내에서의 빈발 프로그램을 선별할 수 있다. 즉, 투사과정을 두 번 거쳐 생성된 빈발 프로그램들로 길이-3의 순차 프로그램 패턴을 추출할 수 있다. 표 2는 투사를 두 번 수행하여 얻어진 길이-3의 순차 프로그램 패턴이다.

표 1. 생성된 요일 별 길이-2의 순차 프로그램 패턴
Table 1. Length-2 sequential patterns

요일	발견된 길이-2의 순차 프로그램 패턴
일	<주말극장-특별기획>, <주말극장-시사매거진2580>, <개그콘서트-타임머신>, <도전1000곡-특별기획>... 총 263개
월	<전파견문록-SBS대하드라마>,<일일연속극-월화드라마>, <TV소설-월화드라마>.. 총 18개
화	<TV특종놀라운세상-월화드라마>, <TV소설-월화드라마>, <일일연속극-월화드라마>.. 총 23개
수	<SBS대기획-섹션TV연예통신>, <일일연속극-드라마스페셜>, <와e멋진세상-일일연속극> .. 총 36개
목	<SBS대기획-생방송한밤의TV연예>, <우리시대-일일연속극>, <피플세상속으로-일일연속극> .. 총 46개
금	<와우동물천하-부부클리닉>, <MBC베스트극장-부부클리닉>, <VJ특공대-부부클리닉>.. 총 34개
토	<그것이알고싶다-터닝포인트>, <주말극장-그것이알고싶다>, <연예가중계-그것이알고싶다>.. 총 160개

본 논문에서는 길이-3인 순차 프로그램 패턴까지 추출하여 프로그램 스케줄 추천에 사용하였다.

표 1과 표 2에서 볼 수 있듯이, 각 요일별로 상당히 많은 수의 패턴들이 생성된 것을 알 수 있다. 이러한 패턴들을

표 2. 생성된 요일 별 길이-3의 순차 프로그램 패턴
Table 2. Length-3 sequential patterns

요일	발견된 길이-3의 순차 프로그램 패턴
일	<주말극장-특별기획-타임머신>, <주말극장-시사매거진2580-드라마시티>, <개그콘서트-특별기획-100인토론>... 총 1743개
월	<전파견문록-일일연속극-월화드라마>,<일일연속극-일일드라마-SBS대하드라마>, <TV소설-전파견문록-월화드라마>.. 총 140개
화	<TV특종놀라운세상-일일연속극-월화드라마>, <TV소설-이것이인생이다-월화드라마>, <일일연속극-일일드라마-월화드라마>.. 총 112개
수	<일일연속극-SBS대기획-섹션TV연예통신>, <일일연속극-드라마스페셜-섹션TV연예통신>, <와e멋진세상-일일연속극-드라마스페셜> .. 총 141개
목	<우리시대-일일연속극-SBS대기획>, <우리시대-드라마스페셜-한밤의TV연예>, <피플세상속으로-일일연속극-해피투게더> .. 총 161개
금	<와우동물천하-신동엽남희석맨II맨-부부클리닉>, <VJ특공대-MBC베스트극장-부부클리닉>, <일일연속극-MBC베스트극장-부부클리닉> 총 116개
토	<주말극장-그것이알고싶다-터닝포인트>, <주말극장-특별기획-그것이알고싶다>, <연예가중계-특별기획-!느낌표>.. 총 909개

개인 사용자에게 추천하기 위해서는 개인의 다양한 선호도 정보들을 고려하여 개인별 추천 스케줄을 형성해야 한다. 이를 위한 사용자 선호도 추출 방법에 대해 다음 절에서 설명하도록 하겠다.

3.4 사용자 선호도 추출

사용자들의 개인별 특성을 고려하기 위해서 본 절에서는 사용자 선호도를 추출하는 방법에 대해 논의한다. 사용자의 선호도에는 다양한 종류의 선호도가 존재할 수 있다. 특정 프로그램에 대한 선호도, 장르에 대한 선호도, 배우에 대한 선호도, 채널에 대한 선호도, 시간대에 대한 선호도 등 방송 프로그램 시청 환경에서 고려할 수 있는 사용자의 선호도는 다양하다. 본 논문에서는 이러한 선호도 중 사용자의 특성을 효과적으로 반영하는 세 가지 선호도에 대해서 계산한다. 먼저 프로그램 선호도는 사용자의 프로그램에 대한 충성도를 나타낸다. 특정 프로그램에 대해 사용자가 얼마나 빠짐없이 시청했는지에 사용자의 프로그램 시청 시간을 기반으로 다음과 같은 수식으로 계산한다.

$$PP_{i,u_a} = \frac{1}{M}\sum_{k=1}^{M}\frac{wt_{i,u_a,k}}{D_i} \tag{11}$$

M은 사용자가 프로그램 p_i를 시청한 총 횟수를 나타내며, k는 p_i를 시청한 각각의 날짜에 해당한다. 따라서 트레이닝 기간 내의 사용자 u_a가 프로그램 p_i를 시청한 모든 날짜에 대하여, p_i의 방영시간 대비 시청 시간을 모두 더한 값을 총 시청 횟수로 나누어서 각 프로그램에 대한 선호도 값을 계산한다.

두 번째 선호도는 사용자의 장르에 대한 선호도이다. 장르 선호도는 사용자의 총 프로그램 시청 시간대비 특정 장르 시청 시간으로 계산한다. 본 논문에서 사용한 사용자 시청 기록 데이터는 총 8가지 장르로 프로그램들을 구분하였고, 이에 따라 8가지 장르에 대한 선호도 값을 계산할 수 있었다. 장르 선호도 계산을 위한 수식은 다음과 같다.

$$GP_{k,u_a} = \frac{\sum_{i}^{|i\in g_k|} wt_{i,u_a,k}}{\sum_{k=1}^{|g_k|}\sum_{i}^{|i\in g_k|} wt_{i,u_a,k}} \tag{12}$$

수식에서 $|g_k|$는 본 논문에서 장르의 가지 수를 나타내며, 각 장르에 포함된 프로그램 p_i들을 시청한 시청 시간을 전체 장르에 대한 시청 시간으로 나누어 계산한다.

세 번째 선호도는 사용자가 시청하는 채널에 대한 선호도이다. 채널 선호도는 장르 선호도와 마찬가지로 사용자의 총 프로그램 시청 시간대비 특정 채널 시청 시간으로 계산한다. 채널은 총 6개로 각 채널에 대한 선호도 값을 다음과 같은 수식으로 계산하였다.

$$CP_{l,u_a} = \frac{\sum_{i}^{|i\in c_l|} wt_{i,u_a,l}}{\sum_{l=1}^{|c_l|}\sum_{i}^{|i\in c_l|} wt_{i,u_a,l}} \tag{13}$$

수식에서 $|c_l|$는 본 논문에서 사용한 채널의 가지 수를 나타내며, 각 채널에 포함된 프로그램 p_i들을 시청한 시청 시간을 전체 채널에 대한 시청 시간으로 나누어 계산한다.

3.5 개인화 프로그램 스케줄 생성

본 절에서는 3.2에서 추출한 유사 사용자 그룹 내의 순차 프로그램 패턴과 3.3에서 추출한 사용자 선호도 값으로 사용자 개인의 프로그램 스케줄을 형성한다. 유사 시청 그룹 내에서 발견된 수많은 패턴들 중에서 개인 사용자에게 적합한 프로그램 패턴을 추천할 수 있도록 사용자의 선호도 값을 적용한다. 다음 식은 개인별 프로그램 패턴을 결정하기 위해 각 프로그램 패턴 별 순위를 계산하는 식이다.

$$rank_{i,u_a} = (1+PP_{i,u_a})*(1+CP_{i,u_a})*(1+GP_{i,u_a})*Q_i \tag{14}$$

이 수식을 이용하여 $rank_{i,u_k}$값을 계산하여 $rank_{i,u_k}$값이 큰 순서대로 각 사용자별 프로그램 패턴의 순위를 매겨 가장 높은 순위를 갖는 프로그램 패턴을 추천한다.

Ⅳ. 실험설계 및 결과

본 논문에서 사용한 실험 데이터는 AC 넬슨 코리아가 수집한 1100명의 2002년 12월부터 2003년 5월까지의 지상파 방송 프로그램 시청 기록이다. 프로그램 시청 기록은 사용자 식별 id, 사용자의 성별, 연령대, 시청 날짜, 시청 시간, 시청 프로그램 제목, 시청 시작 시간, 종료 시간 및 시청 프로그램의 채널, 장르로 구성되어 있다. 이러한 시청 기록 중 2002년 12월부터 2003년 2월까지의 시청 기록을 훈련 데이터로 하였고 2003년 3월부터 2003년 5월까지의 시청 기록을 테스트 데이터로 하여 실험을 설계하였다. 훈련 데이터를 기반으로 30명의 사용자로 구성된 유사 사용자 그룹을 형성하였고 유사 사용자 그룹 내의 순차적인 프로그램 패턴을 발견하였으며, 사용자들의 선호도를 추출하여 사용자 개인의 프로그램 패턴을 추출하였다. 훈련 데이터를 이용하여 생성된 개인 사용자의 프로그램 패턴을 생성하고, 테스트 데이터를 이용하여 생성된 프로그램 패턴에 대해 검증하는 실험을 수행하였다. 실험은 총 세 가지 형태로 수행하였다. 각각의 실험에 대해 다음 절에서 자세히 설명한다.

1. 생성한 패턴들의 추천 성능

3.2절에서 생성한 유사 시청 사용자 그룹의 순차 프로그램 패턴들이 동일한 사용자 그룹의 테스트 데이터에서 실제로 발견되었는지를 확인함으로써 생성한 패턴들의 추천 성능을 계산할 수 있다. 생성한 패턴들은 길이-1, 길이-2, 길이-3인 순차 패턴들로 각 요일별 테스트 데이터를 나누어 패턴들이 존재하는지 확인할 수 있다. 다음 수식을 이용하여 추출된 패턴들에 대한 추천 성능을 계산하였다.

$$\mathrm{Precision}_{L1,d_i,G}=\frac{\sum_{a=1}^{N}\frac{\left|\left\{WL1_{d_i,u_a}:WL1_{d_i,u_a}\in Length1_{d_i}\right\}\right|}{\left|Length1_{d_i}\right|}}{N} \quad (15)$$

$$\mathrm{Precision}_{L2,d_i,G}=\frac{\sum_{a=1}^{N}\frac{\left|\left\{WL2_{d_i,u_a}:WL2_{d_i,u_a}\in Length2_{d_i}\right\}\right|}{\left|Length2_{d_i}\right|}}{N} \quad (16)$$

$$\mathrm{Precision}_{L3,d_i,G}=\frac{\sum_{a=1}^{N}\frac{\left|\left\{WL3_{d_i,u_a}:WL3_{d_i,u_a}\in Length3_{d_i}\right\}\right|}{\left|Length3_{d_i}\right|}}{N} \quad (17)$$

$\mathrm{Precision}_{L1,d_i,G}$는 d_i요일의 길이1인 프로그램 패턴에 대한 유사 시청 사용자 그룹 전체의 정확도 값을 나타낸다. 따라서 분모의 N은 유사 시청 사용자 그룹의 총 사용자 수를 나타내고, $\left|Length1_{d_i}\right|$ 는 d_i요일의 빈발 프로그램들로 선별된 길이-1의 프로그램의 총 수를 나타낸다. 또한 $WL1_{d_i,u_a}$는 사용자 u_a가 시청한 길이-1의 프로그램들을 나타내어 실제 추출된 길이-1의 프로그램 패턴을 사용자u_a가 얼마만큼 시청하였는지를 계산할 수 있도록 하였다. 길이-2, 길이-3 인 프로그램 패턴도 마찬가지 방법으로 계산하였다.

수식에 따라 계산한 정확도 값으로 각 요일의 추천 성능을 나타낸 그래프는 다음과 같다. 그래프에서와 같이 길이-1일 때의 성능이 상대적으로 단기 예측이므로 길이-2, 길이-3일 때의 성능보다 좋게 나온 것을 알 수 있다.

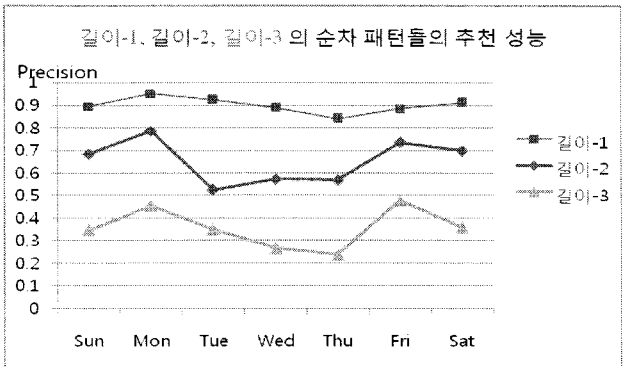

그림 2. 사용자 그룹 전체에 대한 길이1,길이2,길이3의 순차 패턴들의 추천 성능

Fig. 2. Recommendation precisions for recommended sequential TV program of length-1, length-2 and length-3

2. 개인별 프로그램 패턴 추천 성능

본 절에서는 3.5절에서 생성한 개인별 프로그램 패턴 추천 순위에 따라 개인 사용자에 적합한 프로그램 패턴을 추천했을 때의 추천 성능을 계산한다. 즉, 개인 사용자가 실제

로 테스트 데이터에서 그 사용자에 적합한 프로그램 패턴 추천 순위에 따라 시청하였는지 그렇지 않은지를 관찰함으로써 추천 성능을 계산할 수 있다. 다음과 같은 수식을 이용하여 개인별 프로그램 패턴 추천 성능을 계산한다.

$$\begin{aligned} &Precision_{L2,d_i,P} \\ &= \frac{\sum_{a=1}^{N}\sum_{k=1}^{|L1_{d_i,u_a}|}\frac{\left|\left\{WL2_{d_i,l1_k,u_a} : WL2_{d_i,l1_k,u_a} \in PS_{d_i,u_a,l2}\right\}\right|}{\left|\left\{WL2_{d_i,u_a} : WL2_{d_i,u_a} \in Length1_{k,d_i}\right\}\right|}}{N^*\left|L1_{d_i,u_a}\right|} \end{aligned} \quad (18)$$

$$\begin{aligned} &Precision_{L2,d_i,P} \\ &= \frac{\sum_{a=1}^{N}\sum_{k=1}^{|L1_{d_i,u_a}|}\frac{\left|\left\{WL2_{d_i,l1_k,u_a} : WL2_{d_i,l1_k,u_a} \in PS_{d_i,u_a,l2}\right\}\right|}{\left|\left\{WL2_{d_i,u_a} : WL2_{d_i,u_a} \in Length1_{k,d_i}\right\}\right|}}{N^*\left|L1_{d_i,u_a}\right|} \end{aligned} \quad (19)$$

$Precision_{L2,d_i,P}$는 사용자 개개인 별 추천 스케줄에 대한 추천 성능을 측정하기 위한 값으로, d_i요일의 개인별 길이-2인 추천 스케줄의 추천 성능을 뜻한다. 먼저 분모는 사용자의 전체 명수 N과 사용자 u_a가 시청한 길이-1의 프로그램의 총 수를 곱하여 준다. 이는 사용자가 길이-2인 프로그램 스케줄을 구성하는 첫 번째 프로그램이 되는 길이-1의 프로그램을 시청했는지 그렇지 않은지에 따라 그 뒤에 나올 길이-2인 프로그램을 추천하기 때문에 사용자 u_a가 시청한 길이-1의 프로그램 수를 곱해주는 것이다. 분자는 다시 분모와 분자로 나뉘는데, 먼저 분모는 사용자 u_a가 시청한 길이-2의 다양한 프로그램 스케줄 $WL2_{d_i,u_a}$중 길이-1의 특정 프로그램 $Length1_{k,d_i}$을 포함하고 있는 $WL2_{d_i,u_a}$의 수를 나타낸다. 분자는 특정 프로그램 $Length1_{k,d_i}$을 포함하는 사용자 u_a의 길이-2의 시청 프로그램 패턴들 $WL2_{d_i,l1_k,u_a}$중에서 사용자 u_a만을 위한 길이-2의 TV 프로그램 스케줄 $PS_{d_i,u_a,l2}$을 실제로 시청한 횟수를 나타낸다. 즉 훈련 데이터를 이용하여 생성한 사용자 u_a의 길이-2 인 추천 스케줄에 대해 실제로 사용자 u_a가 추천 스케줄의 첫 번째 프로그램을 시청한 후에 두 번째 추천 프로그램인 그 다음 프로그램을 얼마만큼 시청했는지를 계산함으로써 추천 스케줄에 대한 추천 성능을 계산할 수 있었다. 길이-3인 개인 맞춤형 추천 스케줄에 대한 추천 성능도 동일한 방법으로 계산할 수 있다. 수식에 따라 계산한 정확도 값으로 각 요일의 개인별 추천 성능을 그래프로 나타내었다.

그림 4 에서 볼 수 있듯이, 길이-1인 개인별 추천 패턴에 대한 추천 성능이 길이-2, 길이-3보다 더 높은 것을 알 수 있었다. 또한 그룹 전체의 추천 성능보다는 개인별 추천 성능이 약간 더 낮게 측정된 것을 알 수 있었다. 이는 개인의 시청 기록이 일관성 있는 시청 프로그램들로 구성되지 못한 경우에 추천한 프로그램 패턴을 시청하지 않은 경우가 종종 있었기 때문이다.

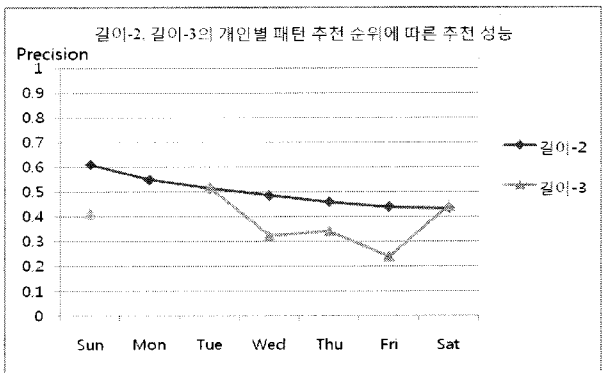

그림 3 길이-2, 길이-3의 개인 맞춤형 추천 스케줄에 대한 추천 성능
Fig. 3. Recommendation precisions of personalized length-2 and length-3 sequential patterns

3. 최소 지지도 값의 변화에 따른 성능 비교

본 절에서 패턴 생성 시 중요한 요소로 작용하는 최소 지지도 값의 변화에 따른 성능에 대해서 실험을 통해 알아보았다. 앞서 논의했듯이 최소 지지도는 패턴의 추출 기준이 되는 값으로 최소 지지도를 넘는 패턴에 대해서만 의미있는 패턴으로 추출한다. 따라서 최소 지지도가 너무 높거나 너무 낮게 되면, 패턴이 거의 생성되지 않거나 또는 너무 많이 생성될 수 있다. 패턴을 생성하는 중요한 기준 값이 되는 최소 지지도 값을 변화시켜 줌에 따라 가장 적절한 패턴을 생성하여 좋은 성능을 나타내는 최소 지지도 값을 결정할 수 있었다. 실험에서 두 가지 최소 지지도 중 발생

빈도(frequency)는 20%로 고정하여 추천 성능을 계산하였다. 성능 측정은 4.1절의 성능 측정 수식과 동일한 수식을 이용하여 계산한다. 성능 값을 그래프로 표현하면 그림 5와 같다.

그림 5 에서와 같이, 최소 지지도를 구성하는 두 가지 요소로 wq값과 발생 빈도(frequency)가 있는데, 이 두 요소의 조합이, wq=3, 발생 빈도(frequency)= 20% 이상일 때에 대해 선별하는 경우가 성능이 가장 좋은 것을 확인할 수 있었다.

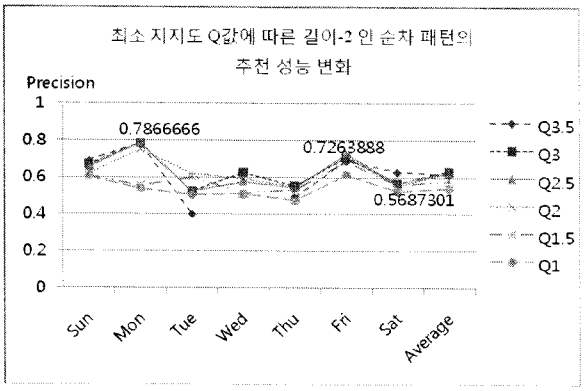

그림 4. 최소 지지도 값에 따른 길이-2 순차 패턴의 추천 성능 변화
Fig. 4. Recommendation precision variation of length-2 sequential patterns with minimum support

4. 기존 PrefixSpan 방법과 제안한 방법과의 성능비교

본 절에서는 기존의 PrefixSpan방법과의 성능 비교를 통해 방송 시청 환경에서 제안한 순차 패턴 마이닝 기법의 우수성을 입증하였다. 기존의 PrefixSpan방법은 패턴 발견 시 아이템의 구매 횟수만을 패턴 선별의 기준으로 삼았다. 이를 방송 프로그램 시청 환경에 적용하면, 아이템 구매 횟수는 단순히 사용자가 아이템인 프로그램을 선택한 횟수를 뜻하며, 프로그램을 선택한 횟수에 따라 패턴을 선별한다는 것이다. 즉, 기존의 PrefixSpan방법 그대로 방송 시청 환경에 적용한 추천 성능과 제안한 순차 패턴 마이닝 기법을 이용한 추천 성능을 동일한 조건에서 실험하였다. 두 가지 실험에 사용한 최소 지지도 값은 다음과 같다.

제안한 알고리즘의 최소지지도: $wq \geq 3, freq. \geq 20\%$
PrefixSpan의 최소 지지도 : $freq. \geq 25\%$

이러한 조건 하에 생성된 패턴들로 추천하였을 때의 추천 성능은 앞서 설명한 4.1절의 성능 측정 방법에 따라 계산하여 그림 6과 같은 결과를 얻었다.

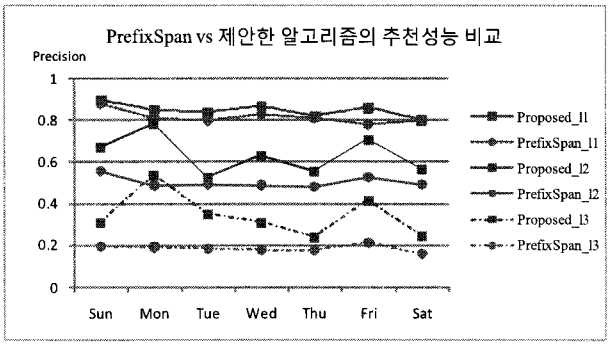

그림 5. PrefixSpan알고리즘과 제안한 알고리즘의 추천 성능 비교
Fig. 5. Comparison of recommendation precisions between our proposed algorithm and original PrefixSpan algorithm

그림 6에서와 같이 제안된 순차 패턴 마이닝 기법을 적용했을 때 기존의 PrefixSpan 방법을 적용했을 때 최대 37% 향상된 성능을 나타내는 것을 확인할 수 있었다.

V. 결론 및 향후 계획

본 논문에서는 TV 프로그램 시청 환경에서의 효율적인 프로그램 스케줄 추천을 위한 추천 시스템을 제안하였다. 프로그램 스케줄은 순차 패턴 마이닝 기법을 적용하여 발견하였으며, 방송 시청 환경에 적합하도록 기존의 순차 패턴 마이닝 기법에 시청 환경에서 고려되어야 할 부가적인 정보(시청 시간, 프로그램 시청 가능 시간)를 추가하여 패턴 추출에 사용하였다. 이러한 추천 시스템을 검증하기 위해 사용자 30명의 약 6개월 간의 시청 기록을 가지고 훈련 데이터, 테스트 데이터로 각각 나누어 추천 성능에 대해 실험하였다. 실험을 통해 제안한 순차 패턴 마이닝 기법이 기존의 순차 패턴 마이닝 기법보다 최대 37% 향상된 성능을

나타냄을 입증하였으며, 추천한 프로그램들이 어느 정도 사용자에게 유용한 추천을 제공한다는 것을 알 수 있었다.

본 논문에서는 실험 데이터의 제약으로 인해 실제 IPTV 시청 환경에서 고려되어야 할 비실시간 콘텐츠인 VoD 콘텐츠, UCC콘텐츠 시청에 대한 부분을 다루지 못하였다. 따라서 향후 계획으로는 실제 IPTV 시청 기록 데이터를 기반으로 지상파 방송과 달리, 콘텐츠의 방영 시작 시간이 정해져 있지 않은 VoD 콘텐츠들에 대한 사용자의 시청 기록에 대해 분석하고, 기존의 지상파 TV 방송 콘텐츠와 VoD 특성을 모두 고려한 순차 패턴 마이닝 기법을 연구할 계획이다. 또한 개인 맞춤형 추천에 대한 사용자의 feedback 정보를 이용하여 보다 적응적인 추천 방법 및 시스템에 대해 연구할 계획이다.

참 고 문 헌

[1] Rakesh Agrawal, Ramakrishnan Srikant, "Mining sequential patterns," 11th International Conference on Data Engineering (ICDE'95), 1995 IEEE Transactions on Software Engineering, pp.3-14, 1995.

[2] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," In Proc. 2001 Int. Conf. Data Engineering, April 2001.

[3] Baoyao Zhou, Siu Cheung Hui, Kuiyu Chang, "An Intelligent Recommender System using Sequential Web Access Patterns," In Proc. 2004 IEEE Conference on Cybernetics and Intelligent Systems, December 2004.

[4] 김성학, 이창훈, "웹 로그 분석을 이용한 추천 에이전트의 개발," 정보과학회논문지 : 기술교육 제2권 제1호, pp.60-66, 2005.

[5] Qiankun Zhao, Sourav S. Bhowmick, "Sequential Pattern Mining: A Survey," Technical Report, CAIS, Nanyang Technological University , Singapore, No. 2003118, pp.1-27, 2003.

[6] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, "Freespan: Frequent pattern-projected sequential pattern mining," In Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), pp.355-359, August 2000.

[7] 강우현, 서용무, 박승봉, "Web Usage Mining을 이용한 온라인 사용자의 웹 탐색패턴 연구," 한국인터넷정보학회 2004 춘계학술발표대회 논문집, pp.99-102, 2004.

[8] R.Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," In Proc. 9th IEEE International Conference of Tools with Artificial Intelligence, 1997

저 자 소 개

표 신 지

- 2007년 2월 : 한국정보통신대학교 전자통신공학과, 공학사
- 2009년 2월 : 한국정보통신대학교 공학부, 석사
- 2009년 3월 ~ 현재 : 한국과학기술원 정보통신공학과 박사과정
- 주관심 분야 : 패턴인식, Personalized IPTV service

김 은 희

- 2000년 2월 : 충남대학교 정보통신공학과 공학사
- 2000년 2월 ~ 2003년 3월 : 삼성전자 Visual Display 사업부 연구원
- 2000년 3월 ~ 2007년 1월 : 삼성전자 Digital Solution Center 선임 연구원
- 2009년 2월 : 한국정보통신대학교, 공학부 석사
- 2009년 3월 ~ 현재 : 한국과학기술원 전기및전자공학과 박사과정
- 주관심 분야 : 패턴인식, Recommendation Agent, IPTV, Object tracking in video

저 자 소 개

김 문 철

- 1989년 2월 : 경북대학교 전자공학과 공학사
- 1992년 12월 : University of Florida, Electrical and Computer Engineering, 석사
- 1996년 8월 : University of Florida, Electrical and Computer Engineering, 박사
- 1997년 1월 ~ 2001년 2월 : 한국전자통신연구원, 선임연구원
- 2001년 2월 ~ 2005년 8월 : 한국정보통신대학교 공학부 조교수
- 2005년 9월 ~ 2009년 2월 : 한국정보통신대학교 공학부 부교수
- 2009년 3월 ~ 현재 : 한국과학기술원 전기및전자공학과/정보통신공학과 부교수
- 주관심분야 : 비디오코딩, 패턴인식, 비주얼 정보처리, UHDTV, IPTV, UXTV, 멀티미디어 시스템