일반논문-09-14-2-14

협업 필터링을 이용한 순위 정렬 모델 기반 (IP)TV 프로그램 자동 추천

김 은 희[a)], 표 신 지[b)], 김 문 철[a)‡]

Automatic Recommendation of (IP)TV programs based on A Rank Model using Collaborative Filtering

Eunhui Kim[a)], ShinJee Pyo[b)], and Munchurl Kim[a)‡]

요 약

방송과 융합의 시대로 접어들면서 (IP)TV 단말에서 이용 가능한 프로그램 콘텐츠 수가 급격히 증가 하였다. 이로 인해, 사용자(시청자)가 선호하는 방송 프로그램 콘텐츠로의 접근성이 주요한 사항이 되었다. 본 논문은 유사 사용자 선호도에 기반을 둔 협업 필터링을 이용하여(IP)TV 프로그램을 효율적으로 사용자에게 자동 추천하는 연구에 관한 내용이다. 개인의 시청 프로그램 선호도를 고려하여 방송 프로그램을 추천하기 위해서, 제안하는 추천 시스템의 구성은 오프라인과 온라인 연산으로 구성된다. 오프라인 연산과정에서 (IP)TV 프로그램, 장르, 채널에 대한 개인의 선호도를 묵시적으로 추론 하는 방법을 제시하고, 동적 퍼지 클러스터링 방법을 사용하여 각 개인의 선호도에 따라 사용자들을 그룹 짓되, 특징 벡터를 장르와 채널에 대한 선호도로 결합하여 사용하는 방법을 제시한다. 또한, (IP)TV 단말에 로그인 한 활동 사용자에게, 높은 정확도로 선호 프로그램을 추천하기 위해서, 활동 사용자와 관심 시청 프로그램이 유사한 사용자들을 유사도 측정 방법을 사용하여 한 번 더 추출하고, 이 추출된 유사 취향 사용자들의 선호 (IP)TV 프로그램들에 대해, EPG를 이용하여 현재 방송되지 않는 프로그램들을 제외시킨다. 마지막 단계에서는 추천 후보 프로그램들에 대해 본 논문에서 제안하는 순위 정렬 모델을 이용하여 추천 우선순위를 결정하여 제시한다. 특별히, 본 논문은 BM(Best Match) 알고리즘을 확장하여 개인 선호도를 고려한 순위 정렬 모델을 제시한다. 실험을 통해, 본 논문에서 제안한 프로그램 자동 추천 알고리듬은 2,441명의 사용자에 대해 5개의 프로그램을 추천하였을 경우, 62.1%의 예측 정확도를 나타내었다.

Abstract

Due to the rapid increase of available contents via the convergence of broadcasting and internet, the efficient access to personally preferred contents has become an important issue. In this paper, for recommendation scheme for TV programs using a collaborative filtering technique is studied. For recommendation of user preferred TV programs, our proposed recommendation scheme consists of offline and online computation. About offline computation, we propose reasoning implicitly each user's preference in TV programs in terms of program contents, genres and channels, and propose clustering users based on each user's preferences in terms of genres and channels by dynamic fuzzy clustering method. After an active user logs in, to recommend TV programs to the user with high accuracy, the online computation includes pulling similar users to an active user by similarity measure based on the standard preference list of active user and filtering-out of the watched TV programs of the similar users, which do not exist in EPG and ranking of the remaining TV programs by proposed rank model. Especially, in this paper, the BM (Best Match) algorithm is extended to make the recommended TV programs be ranked by taking into account user's preferences. The experimental results show that the proposed scheme with the extended BM model yields 62.1% of prediction accuracy in top five recommendations for the TV watching history of 2,441 people.

Keywords : Collaborative Filtering, IPTV, Recommendation System, Rank Model, Dynamic Fuzzy Clustering

Ⅰ. 서 론

방송과 통신의 융합으로 사용자 TV 단말에서 소비 가능한 방송 프로그램의 수가 급격히 증가하였다. 뿐만 아니라, 라이브TV 채널, IPTV 서비스 혹은 TV 포탈과 같은 다양한 콘텐츠 제공 서비스를 통해, 사용자 TV 단말에서 이용 가능한 프로그램 콘텐츠의 양이 기하급수적으로 증대 되었다. 사용자가 자신이 선호하는 방송 프로그램 콘텐츠를 시청하기 위한 방법에는, 프로그램 콘텐츠 검색이나 자동 추천과 같은 형태의 접근 방법이 있다. 그런데, 콘텐츠 검색의 경우, 사용자는 찾고자 하는 프로그램의 키워드를 찾아 검색 엔진에 입력하고, 해당 검색 결과를 기다려야 한다. 또한 검색 결과로 제시된 프로그램 콘텐츠 중에서 자신이 원하는 프로그램 콘텐츠를 선택해야 하는 단계를 거쳐야 한다. 그러나 자동 추천 방법을 사용할 경우, 이러한 여러 단계를 줄일 수 있고, 또한, 협업 필터링을 이용하여 추천을 할 경우, 사용자가 인지하고 있지 못하지만, 선호할 만한 프로그램 콘텐츠를 제시해줄 수 있는 장점이 있다.[1] 이것은 협업 필터링이 유사 선호도 사용자 그룹의 선호 프로그램 콘텐츠를 이용한 추천을 수행하기 때문에 가능하다.

본 논문의 협업 필터링을 이용한 자동 추천 알고리듬은 TV 프로그램의 특성을 고려하여 이루어진다. 제안 시스템은 오프라인과 온라인 계산 과정으로 구성되어 있다.

본 논문에서 개인의 콘텐츠에 대한 선호도 추론은 프로그램 콘텐츠, 장르, 채널에 대한 관심도로 나뉜다. 또한, 협업 필터링을 통해, 각 주어진 활동 사용자에 대해 유사 취향 사용자 그룹을 효율적으로 구성하기 위해서, 두 단계를 거쳐 유사 사용자 클러스터링 과정을 수행한다. 오프라인 계산에서 많은 수의 사용자를 클러스터링 할 때, 조금 더 의미있는 커뮤너티 그룹핑을 할 수 있는 동적 퍼지 클러스터링 방법이 사용된다. 이 때, 유사 사용자의 클러스터링을 위해 사용된 특징 벡터는 장르와 채널에 대한 선호도 값이 사용된다. 이러한 오프라인 계산과정은 시스템의 성능을 고려하여 오버로드가 적은 새벽 시간에 주기적으로 시행됨을 가정한다.

사용자가 (IP)TV 단말에 로그인하면, (IP)TV 프로그램 자동 추천 엔진은 사용자의 프로그램 선호 성향에 맞춘 프로그램을 추천해 준다. 여기서, (IP)TV 단말에 로그인한 사용자를 활동 사용자(active user)라 지칭한다. (IP)TV 프로그램의 추천 단계는 다음과 같다. 먼저 (1) 활동 사용자와 시청 취향이 유사한 사용자들을 Pearson Correlation Coefficient (PCC)기반 유사도 측정 방법을 사용하여 활동 사용자가 속한 클러스터 내에서 찾아낸다. 이 때, 유사도 측정에 사용되는 특징 벡터는 프로그램 콘텐츠에 대한 선호도 값이다. 이후, (2) 유사 취향 사용자들의 선호 프로그램들을 추출한다. 이 프로그램들은 추천 후보 프로그램들이 된다. 그런데 이 추천 후보 프로그램 들 중에서, (3) EPG 정보를 이용해서, 방송 프로그램 편성에서 제공되지 않는 TV 프로그램들을 추천 후보에서 제외시킴으로써 추천 성능을 향상시킨다. 마지막으로, (4) 최종적으로 남은 추천 후보 프로그램들을 협업 필터링에 기반을 둔 순위정렬 모델을 이용하여 추천 순위로 정렬하여 프로그램 콘텐츠를 최종 추천 한다.

본 논문은 다음과 같이 구성되어 있다.

Ⅱ장에서는 먼저 이전에 연구되어 온 협업 필터링을 이용한 추천 연구에 대해 간략히 살펴보고, 제안하는 시스템의 아키텍처를 소개한다. 그리고 Ⅲ장 에서는 본 논문에서 제안하는 새로운 TV 프로그램 자동 추천 알고리듬 중, 개인의 선호도 추론과 유사 취향 사용자들을 동적 클러스터링 수행하는 오프라인 연산과정을 설명한다. 그리고 Ⅳ장에서는 본 논문에서 제안하는 새로운 순위 정렬 모델과 함께 온라인 추천 과정에 대해 설명한다. 실험 결과와 분석은 Ⅴ장에 제시되어 있으며, Ⅵ장에서 본 논문의 결론을 제시 한다.

Ⅱ. 관련 연구 및 시스템 개괄

1. 관련 연구

일반적으로 협업 필터링은 사용자 개인이 별점이나 댓글

a) 한국과학기술원, 전기및전자공학과/정보통신공학과
KAIST, Dept. of Electrical Engineering
‡ 교신저자 : 김문철 (mkim@ee.kaist.ac.kr)
※ 본 연구는 지식경제부 및 정보 통신 연구 진흥원의 IT핵심기술개발 사업의 일환으로 수행하였음. [A1100-0801-3015, Development of Open-IPTV Technologies for Wired and Wireless Networks]

과 같이 해당 구매 품목에 대해 명시적으로 표현한 선호도 값을 이용한다.[2,3,4] 그런데, 현재 TV환경에서 사용자로 하여금 시청한 프로그램 콘텐츠에 대한 선호도를 명시적으로 입력하도록 요구하는 것은 매우 어렵고 사용자를 번거롭게 하여 신뢰할 수 있는 선호도에 대한 데이터 수집이 용이하지 않다. 이는 (IP)TV 단말의 사용자 입력을 위한 인터페이스가 아직 편이성을 제공하지 못하고 있는 것과 연관성이 있다. 본 논문은 명시적인 사용자의 프로그램 선호도 값을 요구하지 않고 이를 묵시적으로 추론하여 계산하는 방법을 이용한다.

전통적으로 협업 필터링은 각 개인 사용자들의 물품 구매 내역을 기반으로, 개인 사용자와 유사한 구매취향을 가진 사용자들을 그룹핑 한 후, 활동 사용자가 구매하지 않은 구매내역들에 대해 유사취향사용자들의 구매 아이템을 정렬하여 추천해주는 형식을 취해 왔다[5,6]. 그런데, 여기서 개인 사용자와 유사한 구매 취향을 지닌 사용자들을 가려낼 때, 시스템에서 관리하는 이용자가 늘어나는 경우에 대비하지 못하는 확장성 문제가 주요 이슈가 되어왔다.

이를 해결하기 위해 G. R. Xue.[7]은 두 번에 걸친 클러스터링 과정을 통하여 복잡도 문제에 대한 해결을 제시하였다. 이를 이용하면 오프라인 클러스터링 과정과 온라인 클러스터링 과정을 나누어서, 온라인 추천 속도를 높일 수 있으며, 또한 개별 활동 사용자에 대해 개인적인 시청 취향 유사도가 더욱 높은 사용자들을 추출 할 수 있는 장점이 있다. 한편, 오프라인 클러스터링 과정 시, 하나의 활동 사용자에 대해 하나의 유사 취향 사용자 그룹(클러스터)의 멤버십을 할당하기 보다는, 여러 클러스터에 속할 수 있게 하는 동적 클러스터링 방법이 추천 성능을 향상 시킬 수 있음을 S. H. Min[8]이 실험을 통해 제시 하였다. 따라서 본 논문에서는 이러한 동적 클러스터링 과정을 오프라인 클러스터링 과정에 적용 한다. 그리고 해당 퍼지 클러스터링의 유효성은 Bezd[14]이 제안한 Partition Coefficient(PC), Partition Entropy Coefficient (PE)값을 사용하여 확인한다.

이전 단락에서 언급된 개인 추천 시스템의 주요 이슈 중에 하나인, 사용자 수 증가에 따른 유사 취향 사용자 추출 계산 증가 문제에 대해, 두 단계에 걸친 클러스터링 방법을 통한 해결 책 외에, Deshpande M.[9]가 제시한 아이템 기반의 추천 시스템이 있다. 이 추천 시스템은 사용자-아이템 매트릭스를 변형하여 아이템-아이템 매트릭스를 형성한 후, 이를 활동 사용자가 구매한 아이템과 곱한 결과를 정렬하여, 활동 사용자에게 추천 하는 방법을 제시하였다. 이러한 시도는 사용자들의 공통 관심 구매 아이템들을 적절히 선택하는 방법으로, 사용자 수의 증가에도 불구하고 계산을 줄일 수 있었고, 추천 성능도 향상시킬 수 있었다는 연구 결과가 보고되었다.[6,9]

웹 검색에서 사용자가 제시한 질의어(query)에 대해, 질의어의 관련성을 비교하여 상관관계가 높은 항목(추천 후보 문서)들을 순서대로 정렬하여 제시하는 방법은, 협업 필터링 시스템에서 사용자가 구매한 항목과 유사성이 높은 아이템들을 정렬하여 추천 아이템으로 제시한다는 면에서 유사성을 지니고 있다. 이에 J. Wang[10]은 Lafferty의 Language Model(LM)을 변형하여, 협업 필터링 기반의 추천 후보 아이템과 사용자간의 상관관계 모델 (순위 정렬 모델)을 제시하였다. 또한, 이 모델은 Deshpande M.[9] 이 제시한 모델에 비해 나은 추천 정확도 성능을 보임을 실험을 통해 입증하였다.[10]

웹 검색에서 널리 사용되는 또 다른 모델로 BM(Best Match) 모델이 있다. BM 모델은 질의어의 단어들과 유사한 주요 용어들을 많이 포함하고 있는 문서들에게 높은 점수를 주어 정렬하는 방법이다. 문서 내의 용어 빈도수 (term frequency in documents – tf)와 질의어 내의 용어 빈도수 (term frequency in query – qtf)를 이용하여 나타낸 스코어 척도(Score Metric)로 모델을 제시하였다[11,12,13].

본 논문에서 제시하는 순위 정렬 모델은 이 BM15 모델을 확장한다. 제시한 순위 정렬 모델의 구성에 대해서는 Ⅳ장 3절에 자세히 설명된다. 그리고 제시 모델과 J. Wang 모델의 성능 비교 측정 및 개념적 차이 분석에 대해서는 Ⅴ장 5절에 설명된다.

2. 개인화 TV 프로그램 자동 추천 시스템의 구성

본 논문에서는 사용자들이 (IP)TV 단말에서 시청한 프로그램들의 시청 히스토리 정보가 서버에서 수집될 수 있도록 IPTV 단말 이전의 TV 단말들은 OCAP, DVB의 out-

of-band 채널을 통한 서버와의 통신을 가정한다.

그림 1은 제안된 협업 필터링 기반의 (IP)TV 프로그램 자동 추천 시스템의 구성 도를 나타낸다. (IP)TV 프로그램 자동 추천 시스템은 오프라인과 온라인 두 계산과정으로 구성되어 있다. 오프라인 계산 과정은 각 개인의 선호 프로그램에 대한 추론이 먼저 이뤄지고, 다수의 사용자들이 주기적으로 동적 퍼지 클러스터링 기법에 의해 방송 프로그램 시청에 대해 유사 취향 사용자 그룹으로 클러스터링 된다.

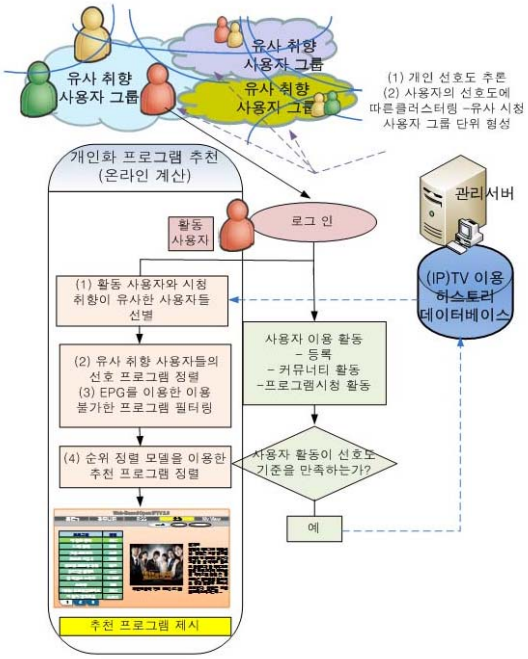


그림 1. (IP)TV 프로그램 자동 추천 시스템 구성도
Fig. 1. Automatic (IP)TV Program Recommendation System Architecture

온라인 계산 과정은, 활동 사용자에게 자신의 선호도에 맞는 TV 프로그램 콘텐츠를 추천해주는 과정이다. 한 사용자가 IPTV 시스템에 로그인 하고 나면 활동 사용자가 된다. 우선, 활동 사용자와 유사한 시청 패턴을 보이는 유사 취향의 사용자들을 활동 사용자가 속한 유사 취향 사용자 그룹들 내에서 선별한다. 그리고 유사 시청 패턴 사용자들의 선호 프로그램들을 추출한다. 이 중, Electronic Program Guide (EPG)를 이용하여 추출된 방송 프로그램들 중에 현재 방송되지 않는 프로그램들을 추천 후보 항목에서 제외한다. 마지막으로 이 추천 후보 프로그램들을 순위 정렬 모델을 이용하여 추천 순위를 정렬한 후, 상위 순위의 몇 개의 프로그램들을 활동 사용자에게 제시함으로써 TV 프로그램 자동 추천 과정을 최종 완료 한다.

III. 오프라인 연산 (선호도 추론 및 클러스터링)

1. 개인 시청 선호도 추론

방송 프로그램 시청 패턴 또는 시청 성향이 비슷한 사용자들을 대상으로 협업 필터링을 수행하기 위해서는 먼저 유사 시청 취향을 갖는 사용자들의 클러스터링이 필요하다. 본 논문에서는 각 프로그램에 대한 선호도를 기반으로, 장르, 채널에 대한 선호도를 계산하고, 이를 기반으로 유사 취향 사용자들을 클러스터링 한다. 먼저 장르와 채널에 대한 선호도 추론을 위해서는 장르와 채널과 같은 메타데이터가 프로그램 콘텐츠 제공 및 소비 시에 이용 가능하여야 한다. 본 논문에서는 각 방송 프로그램 콘텐츠에 대한 메타데이터는 8개중 하나의 장르와, 6개중 하나의 채널로 표 1과 같이 구성되어 있다. 따라서 각 프로그램의 장르와 채널에 대한 사용자들의 선호도를 유사 시청 사용자들을 클러스터링 하기 위한 특징 벡터로 사용한다.

표 1. 장르, 채널 메타데이터 description
Table 1. genre, channel metadata description

	장르(8)	채널(6)
1	교육	ITV
2	드라마 & 영화	KBS2
3	정보	KBS1
4	스포츠	MBC
5	엔터테인먼트	EBS
6	어린이(유아)	SBS
7	뉴스	-
8	기타	-

본 논문에서는 사용자의 (IP)TV 프로그램 콘텐츠에 대한

선호도를 (IP)TV 프로그램 전체 방영횟수 대비 관심 시청 횟수로 정의 한다. 이 때, 관심 시청 대상 프로그램으로 고려되기 위해서는 프로그램 방영 길이 대비, 실제 시청 길이가 전체의 10% 이상인 것을 기준으로 한다. 이는 채널을 변경하는 중에 잠깐 동안 시청한 프로그램들은 관심 프로그램에서 제외함으로써 신뢰도를 증가시키기 위함이다. 그리고 뉴스 및 드라마와 같이 하나의 타이틀로 연속 방영되는 시리즈물 프로그램들은 하나의 프로그램으로 간주한다. 다시 말해, 사용자 u의 프로그램 p 에 대한 프로그램 선호도 x_p^u는 다음과 같이 정의한다.

$$x_p^u = \frac{N_p^u}{N_P} \tag{1}$$

$$N_p^u = \sum_{p'=1}^{N_P} N_{p'}^u \tag{2}$$

$$N_{p'}^u = \begin{cases} 1\,, \text{if}\, \dfrac{W^{u_{p'}}}{L_{p'}} > 0.1 \\ 0\,, otherwise \end{cases} \tag{3}$$

여기서 $N_{p'}^u$는 사용자 u의 한 회분 프로그램 p'에 대한 관심 시청 여부를 0 또는 1로 나타내며, N_p^u는 사용자 u의 개별 프로그램 p'로 이루어진 시리즈물 전체 프로그램 p의 총 관심 시청 횟수를 나타낸다. N_P는 시리즈물 전체 프로그램 p의 총 방영 횟수를 나타낸다. 다시 말해, 한 번의 방영만 한 프로그램의 경우, $N_P = 1$이다. 그리고 $L_{p'}$는 프로그램 p'의 한 회분 방영 길이를 나타내고, $W_{p'}^u$는 사용자 u가 한 회분 프로그램 p'은 실제 시청한 길이를 나타낸다.

장르 선호도는 이 프로그램에 대한 선호도를 이용하여 프로그램이 속한 장르의 선호도 값을 더하여 계산한다. 그리고 이 장르 선호도 값은 모든 장르에 속한 선호도 값을 더한 값으로 나누어 정규화 한다. 전체 장르 개수를 N_G라 할 때, 장르 g에 대한 사용자 u의 선호도 x_g^u는 다음과 같이 정의한다.

$$x_g^u = \frac{\sum_{p \in g} x_p^u}{\sum_{g'=1}^{N_G} x_{g'}^u} \tag{4}$$

장르 선호도를 계산한 것과 같은 방법으로, 전체 채널의 개수를 N_C라 할 때, 사용자 u의 채널 c에 대한 선호도 x_c^u는 다음과 같다.

$$x_c^u = \frac{\sum_{p \in c} x_p^u}{\sum_{c'=1}^{N_C} x_{c'}^u} \tag{5}$$

2. 유사 취향 사용자 클러스터링

많은 수의 사용자들을 클러스터링 할 때, 계산의 효율성을 고려한 합리적인 크기의 단위가 필요하다. 도메인 정보나 위치정보를 사용하여서 사용자들을 일정 수의 기본 단위로 나눌 수 있다. 그리고 이 단위 내에서 개인 선호도 기반 클러스터링 과정을 통해 유사 취향 사용자 그룹들을 나눌 수 있다.

본 논문에서는 2,441명의 전체 Test 사용자들을, 사용자들의 선호도에 따른 유사 취향 사용자 그룹 단위로 나누기 위해, 동적 퍼지 클러스터링(dynamic fuzzy clustering) 방법을 사용하였고, 클러스터의 유효성 확인은 PC(Partition Coefficient) 와 PE(Partition Entropy Coefficient)를[14,17] 사용하였다.

2.1 유사 선호 사용자들에 대한 제안 클러스터링 방법

클러스터링을 위해 널리 사용 되는 기술로 C-means 클러스터링이 있다. 동적 퍼지 클러스터링은 여러 개의 클러스터의 멤버로 속하는 요소가 있을 수 있다는 점에서 C-means 클러스터링과 차이가 난다. 그런데, 이러한 접근은 한 사용자가 하나 이상의 유사 취향 사용자 그룹의 멤버십을 가질 수 있다는 면에서, 실제 응용에 더욱 가까운 클러스터링이라 할 수 있다. Bezd[14]에 의한 퍼지 C-means 클러스

터링 방법 또한 fuzzier 특성과 함께, 그 방법의 단순성으로 널리 이용된다.

퍼지 C-means 클러스터링 과정은 다음과 같이 요약할 수 있다. N이 사용자 전체 수이고, N_K는 각 사용자를 나타내는 특정 벡터의 차원(속성의 개수)이라 할 때, 소속(멤버십)이 결정되어 있지 않은 다음과 같은 $N \times N_K$ 매트릭스가 있다고 가정하자.

$$\left\{x_1^{k'_1},\cdots,x_N^{k'_{N_K}}\right\}=\left\{\begin{array}{c}\left\{\left\{x_1^{g_1},\cdots,x_1^{g_{N_G}}\right\},\left\{x_1^{c_1},\cdots,x_1^{c_{N_C}}\right\}\right\}\\ \cdots\left\{\left\{x_N^{g_1},\cdots,x_N^{g_{N_G}}\right\},\left\{x_N^{c_1},\cdots,x_N^{c_{N_C}}\right\}\right\}\end{array}\right\} \quad (6)$$

여기서 k'은 속성을 대표하는 인덱스로 k'_1은 첫 번째 속성을 나타내고, k'_{N_K}는 N_K번째 속성에 대한 인덱스이다. $x_1^{k_1'}$는 첫 번째 사용자의 k'_1속성에 대한 선호도 값을 의미하고, 마찬가지로, $x_N^{k_{N_K}'}$는 N 번째 사용자의 k'_{N_K}속성에 대한 선호도 값을 의미한다. 그리고 g는 장르 속성을 나타내며 g_1은 첫 번째 장르 인덱스를 그리고 g_{N_G}는 장르 마지막 인덱스를 표시한다. 마찬가지로, c는 채널 속성을 나타내며 c_1은 첫 번째 채널 인덱스를 그리고 c_{N_C}는 채널 마지막 인덱스를 표시한다. 즉, 프로그램이 갖는 N_P개의 속성들, 본 논문에서는 장르, 채널에 대한 선호도가 특징 벡터로 사용되었다. 즉, $N_K = N_G + N_C$.먼저, 퍼지 C-means 알고리듬은 클러스터 개수 $C_s(1 < C_s < N)$와 fuzzifier 값 $m(1 < m)$이 사전에 주어져야 한다. 적절한 C_s와 m값을 찾는 자세한 과정은 다음 (2)항에 설명 되어 있다. 각 클러스터의 중앙값 v_i와 개별 사용자의 클러스터에 대한 멤버십 값 $u_{ij}(C_s \times N)$은 정지 조건이 충족될 때까지 반복해서 계산된다. 정지 조건은 목적 함수 $J_m(U, V)$로 결정된다. 예를 들어, $J_m(U, V)$이전 값과 현재 값의 차이가 기준이 되는 ϵ보다 작은 값이 되면, 반복 계산 과정을 멈춘다.

장르와 채널에 대한 시청 선호도를 특징 벡터로 이용한 제안 퍼지 C-means 클러스터링 과정은 다음 표2와 같다.

본 논문에서 정지 조건은 $\epsilon < 10^{-8}$ 이거나 반복 횟수가 300회를 넘지 않도록 하였다. 그리고 사용자들의 소속 클러스터를 선택하는 방법은, 우선 개인별 멤버십 값의 평균값 $avg_j = \sum_{i=1}^{C_s} u_{ij} / C_s$을 구한 후, 개별 사용자 별 소속 클러스터를 $u_{ij} > avg_j + 0.2$와 같은 기준으로 선택하도록 하였다. 그리고 개별 사용자가 속할 수 있는 클러스터의 최대 개수는 3개로 한정 지었다. 만일, 사용자들이 클러스터들에 더욱 퍼지하게 소속시키려면, 소속 클러스터를 정하는 기준 값을 느슨하게, 예를 들어, 멤버십 평균값보다 같거나 작은 클러스터들을 모두 소속 클러스터로 정함으로써 가능하다.

표 2 제안 퍼지 c-means clustering 과정
Table 2. proposed fuzzy c-means clustering procedure

N사용자, C_s개의 클러스터, m fuzzifier N_P개의 특징에 의한 N_K개의 대표 값들
u_{ij} 멤버십 값 초기화 $u_{ij} \in [0,1], \forall i \in \{1,\cdots,C_s\}, \forall j \in \{1,\cdots,N\}$ $\sum_{i=1}^{C_s} u_{ij} = 1, \forall j \in \{1,\cdots,N\},$ (7) $0 < \sum_{i=1}^{N} u_{ij} < N, \forall i \in \{1,\cdots,C_s\}$
반복 연산 과정 $v_i = \dfrac{\sum_{j=1}^{N}\sum_{p'=1}^{N_P}\sum_{k'=1}^{N_{K'}} u_{ij}^m \cdot x_j^{p'k'}}{\sum_{j=1}^{N} u_{ij}^m}$ (8) $u_{ij} = \dfrac{1}{\sum_{c'=1}^{C_s}\sum_{p'=1}^{N_P}\sum_{k'=1}^{N_{k'}}\left(\dfrac{\lVert x_j^{p'k'} - v_i \rVert}{\lVert x_j^{p'k'} - v_{c'} \rVert}\right)^{\frac{2}{(m-1)}}}$ (9) $J_m(U,V;x)_t = \sum_{i=1}^{C_s}\sum_{j=1}^{N} u_{ij}^m \sum_{p'=1}^{N_P}\sum_{k'=1}^{N_{k'}} \lVert x_j^{p'k'} - v_i \rVert^2$ (10) 여기서, $\text{if } p' = 1, then\, x_j^{p'k'} = x_j^g, N_K' = N_G$ $\text{if } p' = 2, then\, x_j^{p'k'} = x_j^c, N_K' = N_C$
정지 조건 확인 (if $t > 1$) $E(t) = \lvert J_m(U,V;x)_t - J_m(U,V;x)_{t-1} \rvert < \epsilon ?$ (11)

2.2 PC, PE에 의한 클러스터 유효성 확인

Fuzzier m값과 클러스터의 개수 C_s를 효율적으로 찾는 방법으로 Bezd[14]가 제안한 PC, PE값을 사용하였다.

$$PC = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C_s} u_{ij}^2,\ [1/m,\ 1] \tag{12}$$

$$PE = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{C_s} (u_{ij}\log_a u_{ij}),\ [0, \log_a m] \tag{13}$$

PC값이 최소값 $1/m$에 가까울 때, 모든 멤버들의 소속이 하나의 클러스터에 속할 정도로 멤버십 구분이 거의 되지 않은 상태이고, PC 값이 최댓값 1에 가까울 때, 한 멤버가 하나의 클러스터에만 속한 상태(hard clustering)가 된다. PE값의 최솟값과 최댓값에 대응되는 클러스터링 상태는 PC값의 상태와 반대 상태가 된다. 따라서 PC값 변화의 볼록한 부분과, PE값 변화의 오목한 부분이 m과 C_s값의 최적 값이 된다[14].

그림 2에서 확인 할 수 있듯이, 장르와 채널에 대한 선호도를 특징 벡터로 사용할 경우, 10개의 클러스터와 m=1.25의 fuzzifier 값이 주어졌을 때, PC값이 볼록하고 PE값이 오목한 최적의 상태임을 확인할 수 있다. 이 실험을 위해 $m = \{1, 1.25, 1.5, 1.7, 2.0, 3.0\}$과 $C_s = \{3,5,7,10,13,15\}$ 범위의 값을 변화시키며 PC, PE값의 결과 변화를 관찰하였다.

그림 3은 유효성 확인을 위한 PC값(혹은 PE값) 계산 시간에 대한 실험이다. 여기서, PC값을 계산하는 데 소요되는 시간은 클러스터링을 수행하는데 필요한 시간과 비례한다. 그림 3에서 보듯이, 특징 벡터를 엘리먼트 수 1,081개의 프로그램(아이템)으로 했을 경우에 비해, 장르와 채널에 대한 선호도를 이용하여 14개 엘리먼트로 특징 벡터를 구성하였을 경우, 계산 시간이 수 천 배 빠름을 확인 할 수 있다. 따라서 장르, 채널의 선호도를 특징 벡터를 사용하여 클러스터링을 수행 할 경우, 연산 속도를 현저히 향상 시킬 수 있음을 확인할 수 있다.

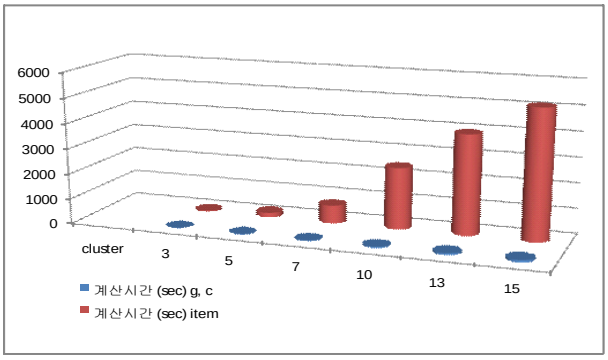
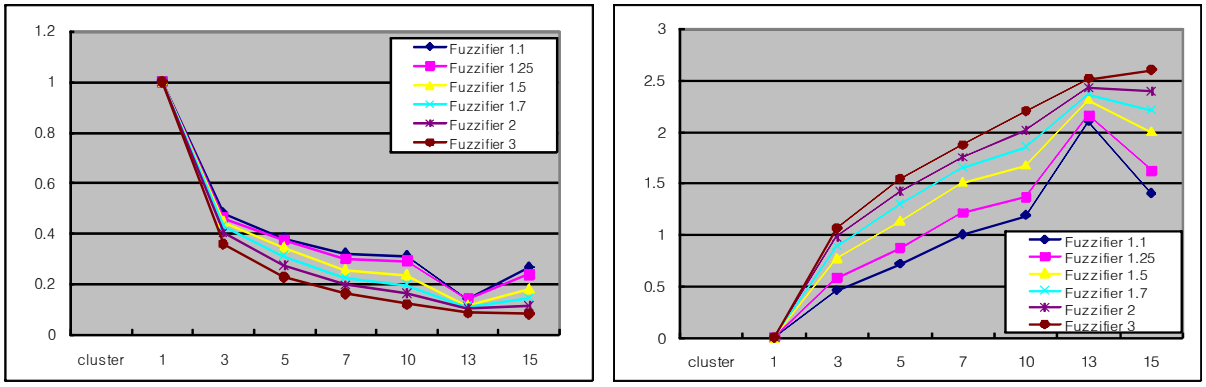


그림 3. PC(혹은 PE) 값 계산 시간
Fig. 3. PC Computation Time

2.3 클러스터의 선호도 특징 분석

이전 항에서 실험을 통해 찾은 각 클러스터 내의 사용자



(a) PC (b) PE

그림 2. PC, PE에 의한 클러스터 유효성 확인
Fig. 2. Cluster Validity Check by PC, PE feature vector - genre, channel

분포는 표 3과 같다. 실험에 사용된 자세한 데이터에 대한 내용은 V장 1절에 설명되어 있다. 전체 사용자 수는 2,441명이며, 마지막 행은 여러 클러스터에 소속한 사용자(Multi joiners)의 수를 나타낸다.

표 3. 각 클러스터 내의 멤버의 수
Table 3. The number of users of clusters

	특징 벡터 - 장르, 채널
클러스터 1	677
클러스터 2	146
클러스터 3	339
클러스터 4	311
클러스터 5	369
클러스터 6	96
클러스터 7	576
클러스터 8	657
클러스터 9	91
클러스터 10	217
다중 클러스터	877

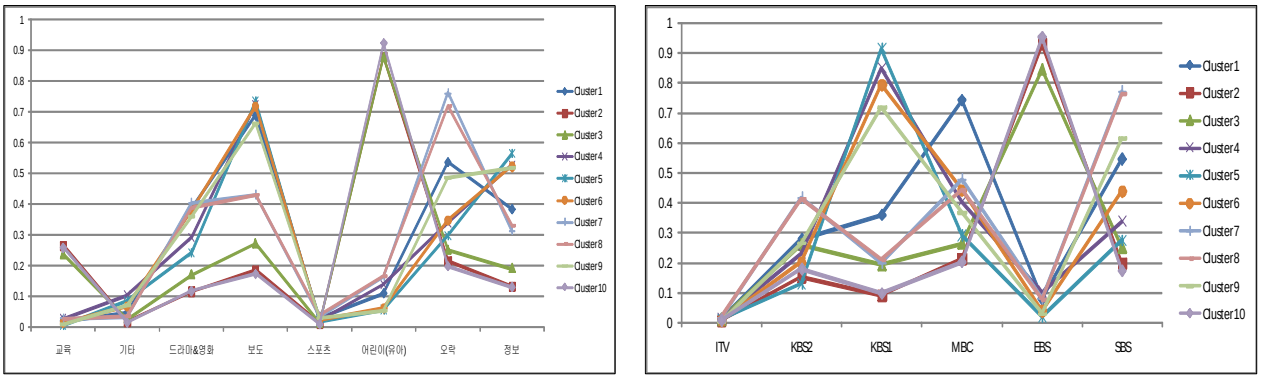
클러스터 각각의 선호도 계산을 위해, 각 클러스터의 장르에 대한 선호도는 해당 클러스터에 속하는 사용자들의 장르에 대한 선호도를 합한 것으로 계산하였다. 그리고 클러스터 내의 사용자 수(N')로 나누어 정규화 하였다. 같은 방법으로 클러스터의 채널에 대한 선호도를 계산하였다. 각 클러스터의 장르에 대한 선호도 y_C^g와 채널에 대한 선호도 y_C^c는 다음과 같다.

$$y_C^g = \frac{\sum_{g'=1, u \in C}^{N_C} x_{g'}^u}{N'} \tag{14}$$

$$y_C^c = \frac{\sum_{c'=1, u \in C}^{N_C} x_{c'}^u}{N'} \tag{15}$$

그림 4에서 보는 바와 같이, 장르와 채널에 대한 클러스터들 간에 독특한 특성들을 발견할 수 있다. 클러스터 1의 경우 오락과 정보 장르를 즐기면서 MBC, SBS, KBS1과 KBS2를 주로 시청하는 것을 알 수 있다. 또한, 어린이(유아)와 교육과 같은 장르를 주로 시청하는 사용자들은 EBS 채널을 시청한다. 클러스터 2, 3, 10이 이와 같은 특성을 보인다. 그리고 클러스터 6의 경우, 보도, 정보, 오락 장르에 관심을 보이면서 KBS1, MBC 그리고 SBS채널을 주요 시청 채널로 하는 것을 확인 할 수 있다. 클러스터 8의 사용자들은 오락과 뉴스 장르에 관심을 갖고 있으며, SBS와 MBC 채널들을 주로 시청하는 것을 확인할 수 있다.

분석 결과에서 보여 주듯이 장르와 채널에 대한 선호도는 서로 밀접히 연관되어 있다.



(a) 10 클러스터의 장르 선호도 (b) 10 클러스터의 채널 선호도

그림 4. 클러스터의 선호도 (FV - g, c)
Fig. 4. Clusters' preferences (FV - g, c)

Ⅳ. 온라인 연산 (개인별 자동 추천 알고리듬)

1. 활동 사용자와의 유사 취향의 사용자 선별

(IP)TV 사용자 단말을 통해 로그인 한 사용자는 활동 사용자가 되고 멤버십을 갖는 유사 취향 사용자 그룹이 결정된다. 해당 유사 취향 사용자 그룹에서 시청 취향이 유사한 사용자를 구체적으로 추출하는 과정을 거치게 된다.

활동 사용자와 유사 취향 사용자 그룹의 사용자간 유사도 측정방법은 PCC (Pearson Correlation Coefficient)를 사용하였다. PCC 는 계산 시 사용자의 선호도 값은 평균 선호도 값으로 감산하여 정규화한 값을 기본 벡터로 사용한다. 따라서 이 방법은 협업 필터링에서 시청 횟수가 많은 사용자의 선호도 값이, 시청 횟수가 적은 사용자의 선호도 값에 비해 크게 반영되는 것을 방지할 수 있어, 더욱 공정한 유사도 측정 방법임이 Resnick P.[5]에 의해 입증되었다.

활동 사용자가 시청한 프로그램 콘텐츠의 리스트를 활동 사용자 u_a의 선호도 프로파일 L_{u_a}로 정의 한다. L_{u_a}에 대해 활동사용자의 선호도 값과 추출된 유사 취향 사용자 그룹의 각 멤버간의 선호도에 대한 유사도를 계산한다. 즉, 활동 사용자 u_a와 사용자 u간의 유사도($\sim(u_a,u)$)는 다음과 같이 측정된다.

$$\sim(u_a,u)=\frac{\sum_{i_m\in L_{u_a}}\left(x_{i_m}^{u}-\overline{x^{u}}\right)\bullet\left(x_{i_m}^{u_a}-\overline{x^{u_a}}\right)}{\sqrt{\sum_{i_m\in L_{u_a}}\left(x_{i_m}^{u}-\overline{x^{u}}\right)^2}\sqrt{\sum_{i_m\in L_{u_a}}\left(x_{i_m}^{u_a}-\overline{x^{u_a}}\right)^2}} \quad (16)$$

여기서 i_m은 활동 사용자가 관심을 보인 시청 프로그램들을 나타내며, $x_{i_m}^{u}$은 사용자 u의 프로그램 i_m에 대한 선호도 값이고, $\overline{x^{u}}$은 사용자 u의 평균 시청 선호도이다. 그리고 L_{u_a}는 활동 사용자 u_a가 시청한 관심 프로그램 집합을 나타낸다.

유사도 값이 0이상인 사용자들을 유사 취향의 사용자로 분류한다. 그리고 추천의 속도를 고려하여, 추천 사용자의 수는 상위 M명으로 제한 할 수 있다. 본 논문에서는 70명으로 한정 지어 실험을 진행하였다.

2. EPG 필터링 과정을 통한 추천 후보 아이템 정렬

활동 사용자와 시청 취향이 유사한 사용자들을 선별한 후, 이들의 선호 시청 프로그램들은 추천을 위한 후보 프로그램이 된다. 실시간 방송 프로그램을 추천할 경우, EPG 정보를 통해 접근 가능한 프로그램만을 선별할 수 있다. IPTV와 같이 VOD 형태의 프로그램 콘텐츠를 포함하는 경우, EPG는 최근 프로그램이 어떤 것인 지에 대한 정보로 사용 가능하다. 그리고 개인 방송과 같은 형태의 프로그램 제공의 경우[15], 사용자들이 개인 방송 스케줄을 미리 제시하면, 이를 EPG정보로 활용하여 이용 가능한 프로그램들을 추천후보로 선정하는데 사용 될 수 있다.

본 논문에서는 실시간 방송 시청에 대한 히스토리 데이터에 대해 실험을 수행하였다. 실험 결과, 추천 프로그램에서 EPG를 이용해 방송되지 않은 프로그램을 제거함으로써 추천 정확도 성능이 향상됨을 확인할 수 있었다.

3. TV 프로그램 정렬을 위한 BM15 모델의 확장

Ⅱ장 1절에서 BM 모델에 대한 간략한 소개가 있었다. BMX 모델의 대한 자세한 내용은 Robertson SE.[11,12,13,16]의 논문에 설명되어 있다. BM15모델은 다음과 같은 과정을 통해, 본 논문에서 제시하는 순위정렬 모델로 확장되었다. 먼저 검색 결과 후보 문서 d와 질의어 Q의 관계를 추천 후보 프로그램 i_c와 활동 사용자 u_a와의 관계로 재해석 한다. 그리고 활동 사용자 u_a를 활동사용자가 선호하는 프로그램의 프로파일 L_{u_a}로 대체함으로써, 추천 후보 프로그램 i_c와 활동 사용자 시청 선호 프로그램L_{u_a}간의 관계는 다음 식으로 정의된다.

$$Score\left(i_c,L_{u_a}\right)=\sum_{i_b\in L_{u_a}}w^{(1)}w^{(2)}\frac{\left(k_1+1\right)wf\left(i_b,i_c;L_{u_a},U\right)}{k_1+wf\left(i_b,i_c;L_{u_a},U\right)}\frac{\left(k_3+1\right)wf\left(i_b;L_{u_a}\right)}{k_3+wf\left(i_b;L_{u_a}\right)} \quad (17)$$

여기서 $w^{(1)} = \log\frac{(N-n+0.5)}{(n+0.5)}, w^{(2)} = \sim(i_b, i_c)$.

N는 i_c프로그램들의 전체 방영 횟수를, 그리고 n은 i_b 프로그램의 방영 횟수를 나타낸다. i_c는 앞의 2절에서 설명된 바와 같이 추출된 후보 프로그램들을 나타내며, i_b는 L_{u_a}에 속하는 활동 사용자의 선호 프로그램을 의미한다. $wf(i_b, i_c; L_{u_a, U})$는 유사취향 사용자들(U)의 전체 프로그램(NTP)시청 횟수 대비, i_b와i_c 두 프로그램을 시청한 사용자들의 시청 횟수의 합의 비율을 나타낸다. 이 때, 유사 취향 사용자들의 명수를 N_U라 하고,두 프로그램을 시청한 경우를 $c(i_b, i_c) > 0$ 로 표기하면,

$$wf(i_b, i_c; L_{u_a, U}) = \frac{\sum_{u'=1}^{N_U}\left(N_{i_b}^{u'} + N_{i_c}^{u'}\right)\Big|_{c(i_b \cap i_c)>0}}{\sum_{u'=1}^{N_U}\sum_{p'=1}^{N_{TP}} N_{p'}^{u'}} \tag{18}$$

$wf(i_b; L_{u_a})$는 활동 사용자의 프로그램 $N_P = count(L_{u_a})$시청 횟수 대비, i_b프로그램을 시청한 횟수의 비율을 의미한다. 즉,

$$wf(i_b; L_{u_a}) = \frac{N_{i_b}^{u_a}}{\sum_{p'=1}^{N_P} N_{p'}^{u_a}} \tag{19}$$

$w^{(2)}$는 BM15 모델에 추가된 두 아이템 i_b와 i_c간의 유사도에 대한 가중치로, 사용자-아이템(User-Item) 매트릭스에서VCC(Vector Cosine Correlation) 방법으로 식 (20)과 같이 측정된다[9]. 즉,

$$w^{(2)} = \sim(i_b, i_c) = \frac{\overrightarrow{x_{i_b}^*} \bullet \overrightarrow{x_{i_c}^*}}{\| x_{i_b}^* \| \bullet \| x_{i_c}^* \|} \tag{20}$$

그림 5는 두 아이템간의 VCC 유사도 측정의 예를 나타낸다. 그림 5의 사용자-아이템(User-Item) 매트릭스에서, 두 아이템 I3과 I4를 동시에 시청한 두 사용자 P와 T를 볼 수 있다. 여기서 I3과 I4의 유사도 측정치는 0.643으로 평균 유사도 값 0.5 이상의 값이다. 반면에, I1과 I4 프로그램 역시 두 사용자 P와 T가 동시에 시청하였다. 그런데, I1과 I4 아이템간의 유사도 값은 0.4로 평균값 0.5 이하의 값을 갖는다. 따라서 두 프로그램 간의 유사 여부를 판단하는 기준으로 유사도 임계값 0.5를 사용할 경우, I3과 I4의 유사도는 0.643으로 두 프로그램 간 유사도가 있는 것으로 판단이 되고, I1과 I4의 유사도는 0.4로 두 프로그램간의 유사도는 없는 것으로 판정될 수 있다. 따라서 식 (17)에서 $w^{(2)}$는 사용자와 아이템간의 상관도를 보다 의미 있게 예측하는 효과가 있다.

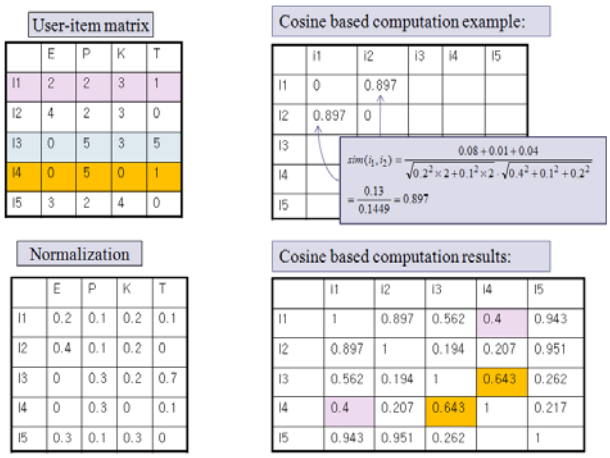


그림 5. 식 (17)에서 가중치 $w^{(2)}$ 역할의 중요성

Fig. 5. The importance of weight $w^{(2)}$ in Eq. (17)

또한 식 (17)은 다음과 같이 직관적으로 해석될 수 있다. 첫 번째 항 $wf(i_b, i_c; L_{u_a, U})$부분은 TV 프로그램에 대한 유사 사용자들의 선호도를 고려한 부분이며, 두 번째 항 $wf(i_b; L_{u_a})$은 활동사용자의 시청 프로그램에 대한 선호도 값을 나타낸다. 그리고 계산식에서 두 항이 곱으로 계산되므로, 유사 사용자들의 선호도와 활동 사용자의 선호도가 일치할 경우, 높은 순위가 되어 추천 상위 항목으로 추천되는 효과가 있기 때문에, 유사 사용자들의 선호도와 활동 사용자의 선호도가 서로 보완 관계에 있음을 알 수 있다.

V. 실 험

1. 실험 데이터

실험 데이터는 2002-12-01부터 2003-05-31, 6개월 동안 2,441명의 사용자의 Nelson Korea의 TV 시청 히스토리 데이터를 사용하였다. 표 1에서처럼, 전체 시청 채널은 6 채널과 8개의 장르로 나뉘어져 있다. 6개월 동안, 전체 시청 프로그램의 타이틀은 1,081개이고, 서브타이틀까지 고려하면, 1,855개이다. 표 4는 TV 시청 히스토리 데이터베이스의 필드에 대해 설명하고 있다. 그리고 실험을 위해서, 훈련(training) 데이터는 2002-12-01부터 2003-03-31까지 4개월을 사용하였으며, 2003-04-01부터 2003-05-31까지 2개월은 검증(test) 데이터로 사용하였다.

표 4. TV 시청 히스토리 데이터베이스 필드
Table 4. Database Fields of TV Usage History

데이터베이스 필드	콘텐츠
ID	사용자의 ID 값
Date	방송일
DateOfWeek	요일
Title	프로그램 타이틀
SubTitle	프로그램 서브타이틀
w_start_t	시청 시작 시간
w_end_t	시청 종료 시간
p_start_t	프로그램 시작 시간
p_end_t	프로그램 종료 시간
Channel	채널
Genre	장르
N_Cluster	User가 속한 클러스터 개수
Cluster1	첫 번째 멤버십 클러스터 번호
Cluster2	두 번째 멤버십 클러스터 번호
Cluster3	세 번째 멤버십 클러스터 번호

2. 실험 성능의 측정

실험 성능은 Precision을 이용하여 측정하였다. 보통 정보 검색분야에서 Precision과 Recall이 성능 측정 방법으로 사용된다. Precision은 추천한 아이템 목록 중에서, 시청자가 실제로 관심 있게 시청한 아이템 항목의 비율이 얼마나 되는지를 나타낸다. 그리고 Recall은 시청자가 실제로 관심 있게 시청한 아이템 중 추천 항목 안에 들어가 있는 아이템의 비율이 얼마나 되는지를 나타낸다. 여기서, 관심 있게 시청한 아이템의 기준은 프로그램의 길이 대비 10% 이상의 시청 여부이다. 이 기준은 Training 데이터 추출 및 Test 데이터 성능 비교에 적용되었다. 그런데, TV 프로그램 추천에 있어서, Precision은 Recall에 비해 더 중요한 의미를 가진다. 왜냐하면, Recall의 정확도는 추천하는 프로그램의 개수가 늘어날 경우 증가 되는데, 더 많은 프로그램을 추천한다는 의미는, 활동 사용자에게 선호 항목을 더욱 많이 제시하게 되므로, 찾아 선택해야 하는 항목이 늘어나므로 불편을 주게 되기 때문이다.

따라서 본 논문에서는 최대 40개까지의 프로그램 추천에 대한 precision의 정확도에 중점을 두도록 하였다.

3. 개인 선호 프로파일과 추천 프로그램 예제

그림 1에서 보여주는 것과 같이, 사용자에게 TV 프로그램을 추천하기 위해서는 각 사용자의 선호도를 먼저 추론하는 과정이 필요하다. 표 5는 ID 103461201의 사용자의 선호도 값 x_p^u에 따라 정렬된 선호 프로그램의 예시를 보여주고 있다. 표 5에는 ID 103461201 사용자의 선호 프로그램 전체 67개 항목 중 10개의 항목이 제시되어 있다.

표 5. ID 103461201 사용자 개인 선호 파일 예시
Table 5. Personal preference profile example of a user with ID 103461201

아이템 인덱스	방영횟 수	시청 횟수	x_p^u	제목
1056	1	1	1.000	TV동물농장
85	4	4	1.000	그것이알고싶다
780	2	2	1.000	와우동물천하
575	1	1	1.000	인생대역전
584	3	3	1.000	일요일일요일밤에
716	1	1	1.000	터닝포인트
713	4	3	0.750	타임머신
915	3	2	0.667	인간극장
748	3	2	0.667	특별기획
556	3	2	0.667	위기탈출수호천사

표 6은 ID 103461201을 지닌 사용자의 추천 항목 10개를 예시로 나타내었다. 이 항목들은 식 (17)의 결과 점수값에 의해 정렬되었다.

표 6. ID 103461201 사용자 추천 예시
Table 6. Recommendation Example of a user with ID 103461201

아이템 인덱스	점수	제목	채널
640	0.015	주말연속극	KBS2
957	0.008	KBS 9시 뉴스	KBS1
593	0.008	일일 연속극	MBC
592	0.008	일일 연속극	KBS1
1000	0.007	MBC 뉴스	MBC
590	0.006	일일 시트콤	MBC
571	0.006	인각 극장	KBS2
1002	0.006	MBC News Desk	MBC
34	0.005	개그 콘서트	KBS2
285	0.004	미니 시리즈	MBC

4. 제안 추천 모델 최적 성능 변수(k_1과 k_3)값

식 (17)에서 k_1과 k_3는 순위 정렬 모델의 tf와 qtf의 비율을 조절하기 위한 변수이다.

qtf는 활동 사용자의 전체 시청 횟수 대비 i_b 아이템 시청 횟수 비율을 의미한다. 식 (17)에서 $(k_3+1)wf(i_b;L_{u_a})/k_3+wf(i_b;L_{u_a})$ 부분은 i_c아이템과 무관한 부분이다. 따라서 k_3는 순위에 영향을 미치지 못한다.

제안 모델은 추천 후보 아이템 $i_c \in L_{u_a,U}$를 점수에 따라 정렬하므로, precision 성능에 영향을 미친다. 실험 결과는 표 7과 같다.

표 7. k1 변수에 따른 Precision 성능
Table 7. Precision Accuracy of k1 variable

변수		Precision 성능				
k3	k1	TOP5	TOP10	TOP20	TOP30	TOP40
200	0.2	0.651	0.642	0.616	0.584	0.556
200	0.6	0.644	0.642	0.614	0.580	0.557
200	1.0	0.642	0.643	0.613	0.580	0.557
200	1.2	0.642	0.642	0.612	0.580	0.557
200	1.5	0.642	0.641	0.612	0.580	0.556
200	2.0	0.642	0.641	0.612	0.580	0.556
200	2.4	0.642	0.641	0.612	0.580	0.556
200	3.6	0.642	0.641	0.612	0.580	0.556

위 실험은 6번 클러스터 내의 93명 사용자들의 precision 성능의 평균값으로 결과를 나타내었다. 93 명의 사용자 중, 한 달의 Test 기간 동안 관심 시청 횟수가 한번 도 없는 2명의 사용자는 추천 성능에서 제외 되었다. 표 7에서 보는 바와 같이, Precision 성능은 Top N개의 아이템 - Top5, Top10, Top20, Top30 그리고 Top40에 대해 각각 측정되었다. 실험을 통해 최적의 Precision 성능은 $k_1=0.2$과 $k_3=200$일 때임을 확인할 수 있다.

5. 제안 추천 모델의 성능 비교

그림 5에 설명되었듯이, VCC (식 20)척도가 두 아이템간의 유사도 측정 가중치$w^{(2)}$로 사용되었다. 이 $w^{(2)}$가 추가된 경우와 그렇지 않은 경우에 대한 비교 실험을 표 8에 제시 하였다. 표8에서 보는바와 같이,$w^{(2)}$ 추가 시, Precision 정확도가 향상되는 것을 확인할 수 있다. 아래 실험 또한 6번 클러스터 사용자들의 Precision 평균값을 이용한 실험 결과이다.

표 8. 가중치 $w^{(2)}$있는 경우와 없는 경우의 성능 비교
Table 8. Precision Performance Comparison with and without $w^{(2)}$

$w^{(2)}$	Precision 성능				
	TOP5	TOP10	TOP20	TOP30	TOP40
VCC	0.723	0.686	0.640	0.602	0.573
None	0.648	0.611	0.580	0.554	0.535

J. Wang이 제시한 순위 정렬 모델과 본 논문을 통해 제시된 순위 정렬 모델의 성능 비교를 위해, 그림 1에 제시된 똑같은 과정을 거친 후, 순위 정렬 모델만을 차이를 둔 비교 실험을 진행하였다. 표 9의 결과와 같이, 본 논문을 통해

표 9. J. Wang 모델과 제시 모델간의 Precision 성능 비교
Table 9. Precision Accuracy Comparison between J. Wang model and proposed model

model	Precision 성능				
	TOP5	TOP10	TOP20	TOP30	TOP40
Propose	0.723	0.686	0.640	0.602	0.573
J. Wang	0.615	0.592	0.563	0.541	0.516

제시된 모델이 J. Wang의 모델에 비해 더 나은 성능을 보임을 확인할 수 있다.

두 모델간의 이론적인 차이에 대해 다음과 같은 분석을 할 수 있다.

J. Wang[10] 은 순위 정렬 모델에서 추천 후보 문서 내에 질의어가 존재하는 경우에 대해서만 고려하였다. 즉, 연산 과정 중 두 프로그램을 시청한 사용자 수에 대해서만 고려하였다. 이에 반해, 제안된 순위 정렬 모델에의 $w^{(1)}$은 추천 후보 문서 내에 질의어가 존재하는 경우와 존재하지 않는 경우 모두에 대해 고려하였다. 자세한 내용은 Robertson S.E.[16]의 논문에 잘 설명되어 있다.

TV 프로그램 추천에 있어서, 유사취향 사용자들의 선호도뿐만 아니라, 활동 사용자의 선호도도 중요하다. 그리고 시리즈인 TV 프로그램의 경우, 대부분의 사용자가 이전 시청 다음 회수 프로그램을 보는 경우가 많으므로, 한 프로그램을 반복 시청하게 된다. 본 논문을 통해 제시한 순위 정렬 모델은 유사 취향 사용자들의 프로그램에 대한 선호도와 활동사용자의 선호도 값을 곱하여 계산함으로써, 서로의 선호도가 일치할 때 높은 순위 결과를 내게 된다. 반면, J. Wang의 모델은 활동사용자의 선호도가 크게 될 경우, 순위 값을 낮추는 결과를 초래한다. 따라서 본 논문이 제시한 순위 정렬 모델이 TV 프로그램 추천에 있어 더 적절하다.

또한 J. Wang 모델에서 조건부 확률 $P(i_b|i_c,r)$은 협업 필터링을 적용하여 개별 사용자들의 시청 횟수에 대한 고려 없이 i_c프로그램 시청 사용자수 대비 두 i_c와 i_b 프로그램을 시청한 사용자 수로 계산하였다. 하지만, 제시한 모델에서는 $P(i_b|i_c,r)$부분을 i_c프로그램을 시청한 사용자들의 총 시청 횟수 합 대비 두 i_c와 i_b 프로그램을 시청한 사용자들의 시청 횟수 합의 비율로 사용함으로써(식 18) 조금 더 정교한 계산을 수행하였다.

마지막으로, 두 i_c와 i_b 프로그램간의 유사도 측정을 추가한 $w^{(2)}$을 통해 정확도를 향상 시킨 모델을 제시할 수 있었다.

6. 개인 선호 프로파일과 추천 프로그램 예제

Ⅲ장 2절 2항에서 설명되었듯이, 전체 2,441명의 Test 사용자에 대해, 장르와 채널을 특징벡터로 동적 퍼지 클러스터링 수행 후, 10개의 클러스터(유사 사용자 그룹)를 얻었다.

표 10에 나타나듯이 서로 다른 클러스터들이 서로 다른 Precision 정확도를 보이는 것을 확인할 수 있다.

Precision 성능에 영향을 미치는 요소로, 활동 사용자와 유사취향의 사용자들이 잘 선별되었는지의 여부와, 활동 사용자가 일정한 장르와 채널에 대한 꾸준한 관심을 유지하는 지 등이 있다.

표 10. 10개 클러스터의 선호 프로그램 추천 정확도 성능
Table 10. Program Recommendation Precision Performance of 10 clusters

* (1)	** (2)	*** (3)	Precision				
			TOP5	TOP10	TOP20	TOP30	TOP40
1	638	63	0.651	0.622	0.587	0.553	0.521
2	142	18	0.573	0.534	0.473	0.450	0.433
3	324	28	0.576	0.528	0.485	0.460	0.438
4	295	16	0.609	0.582	0.524	0.492	0.471
5	358	30	0.587	0.541	0.509	0.482	0.461
6	93	2	0.723	0.686	0.640	0.602	0.573
7	545	64	0.618	0.607	0.575	0.534	0.498
8	624	85	0.600	0.592	0.556	0.519	0.485
9	87	6	0.671	0.652	0.582	0.528	0.489
10	212	18	0.603	0.546	0.506	0.488	0.472
계	2,441		0.621	0.589	0.544	0.511	0.484

* (1) N 번째 클러스터, ** (2) 전체 사용자 수,
*** (3) Test 기간 동안 관심 시청 프로그램이 없는 사용자의 수

실험 결과, 가장 Precision 성능이 우수한 클러스터는 6번 클러스터이고, 가장 Precision 성능이 떨어지는 클러스터는 2번 클러스터였다. 그림 4를 통해 각 클러스터의 특성을 비교해 보면, 2번 클러스터의 사용자들은 EBS채널을 통한 교육과 어린이(유아) 장르의 콘텐츠를 주로 시청하였다. 반면, 6번 클러스터의 사용자들은 KBS1, MBC, SBS 채널들을 통해 보도, 오락 및 정보 장르를 주로 시청하였다. 그리고 2번 클러스터 사용자들의 평균 시청 프로그램의 수는

Training 기간 동안 214개이고, Test 기간 동안 79개이다. 반면, 6번 클러스터 사용자들의 평균 시청 프로그램의 수는 Training 기간 동안 266개이고, Test 기간 동안 115개이다. 이 분석을 통해, 추천 성능이 좋은 사용자들의 그룹은 일반적으로 대중이 선호하는 프로그램들을 시청하는 사용자들이면서 프로그램 시청을 많이 하는 사용자들임을 알 수 있다.

표 9의 마지막 행은 2,441 명 사용자들의 전체 평균 Precision 정확도를 보여준다.

Ⅵ. 결 론

본 논문에서는 동적 퍼지 클러스터링 방법을 통해 하나 이상의 관심 분야를 가진 사용자들이 여러 유사취향사용자 그룹에 속할 수 있는 그룹핑 방법을 이용하였으며, 이전 논문과 다르게 TV 프로그램의 장르와 채널에 대한 선호도와 같은 여러 선호도를 결합한 퍼지 c-means 클러스터링 방법을 제시하였다. 또한, 활동 사용자에게 추천을 해주는 순위 정렬모델은 기존의 협업 필터링에서 사용한 유사 취향의 사용자들의 프로그램 선호도만을 고려했던 것에 반해, 활동사용자의 프로그램 선호도와 프로그램간의 유사도 측정 가중치를 추가하는 모델을 확장 제시하였다.

제안한 협업 필터링을 이용한 순위 정렬 모델 기반 TV 프로그램 자동 추천은 이전 방법들에 비해 더 나은 예측 성능을 확인할 수 있었다. 2,441명의 사용자들에게 5개의 추천 아이템을 제시하였을 경우 62.1%의 Precision 정확도를 보임을 확인할 수 있었다.

실험에 사용된 데이터는 실시간 방송에 대한 시청 히스토리를 기반으로 하였다. 향후, VOD 콘텐츠를 보유한 IPTV에 대한 시청 히스토리 데이터에 대한 추가 실험을 수행할 계획이다. 향후 연구로서는, 협업 필터링의 단점들을 보완하기 위해 콘텐츠 기반 추천방법을 결합한 하이브리드 추천 모델 연구를 진행할 예정이다.

참 고 문 헌

[1] M. MONTANER, B. LÓPEZ, J. L. DE LAROSA, A Taxonomy of Recommender Agents on the Internet, Artificial Intelligence Review 19: 285 - 330, 2003.

[2] Robertson S.E., Sparck Jones K., Relevance Weighting of Search Term, Journal of the America Society for Information Science, May-June, 1976.

[3] L. Greg , S. Brent , and Y. Jeremy. Amazon.com Recommendations, Item-to-Item Collaborative filtering, IEEE Internet Computing, 2003

[4] Cotter, P., and Smyth, B., PTV: Personalised TV Guides. In Proceedings of the 12th Conference on Innovative Applications of Artificial Intelligence, IAAI 2000, Austin, Texas, 2000.

[5] Resnick P., Iacovou N., Suchak M., Bergstrom P., and Riedl. J., “An Open Architecture for Collaborative Filtering of Netnews”, In Proceedings of the ACM Conference on Computer Supported Cooperative Work, (pp 175-186), 1994.

[6] Breese, J. Heckerman, D., and Kadie, C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI., 1998.

[7] Xue G., Lin C., Yang Q., Xi W., Zeng H., Yu Y., Chen Z., “Scalable collaborative filtering using cluster-based smoothing”, International ACM SIGIR conferenceon research developmenton information ,2005.

[8] Min SH., Han I., “Dynamic Fuzzy Clustering for Recommender Systems”, Springer-Verlag Berlin Heidelberg 2005.

[9] Deshpande M., Karvpis G., “Item-based top-N recommendation algorithms”, ACMTrans. Inf. Sys., 2004.

[10] Wang, J., Powelse, J., Fokker, J., Vreies, A., Reinders M., “Personalization on a peer-to-peer television system”, Multimedia Tools and Applications. Vol.36, no1-2, 89-103, Jan. 2007.

[11] Robertson SE., Walker S., Beaulieu M., Gatford M., and Payne A., Okapi at TREC-4. In NIST Special Publication 500-236:The Forth Text Retreival Conference (TREC-4), pages 73-96, Gaithersburg, MD, 1995.

[12] Robertson S.E., Sparck Jones K., Relevance Weighting of Search Term, Journal of the America Society for Information Science, May-June, 1976.

[13] Robertson S.E., Walker S., “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”, In W.B. Croft & C.J. van Rijsbergen (Eds.), SIGIR 94. Proceedings of the 17th International Conference on Research and Development in Information Retrieval, Dublin (pp. 232-241). London: Springer-Verlag, 1994.

[14] Sergios T., Konstantions K. Pattern Recognition, Third Edition. Academic Press, USA, 2006.

[15] Afreeca TV, http://www.afreeca.com

[16] Robertson S.E. and Spark Jones K., “Relevance weighting of search terms” Journal of the American Society for Information Science, 27:129-146, 1976.

[17] J.A. Hartigan, ClusteringAlgorithms.Wiley, 1975.

저 자 소 개

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, 전자공학과, 석사
- 2009년 3월 ~ 현재 : 한국과학기술원 정보통신공학과 박사과정
- 주관심 분야 : 패턴인식, Recommendation Agent, IPTV, Object tracking in video

표 신 지

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

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