IDM을 기반으로 한 사용자 프로파일 예측 및 개인화 추천 기법

*염선희
삼성전자 주식회사, 디지털미디어연구소
e-mail : *sunny.youm@samsung.com*

User Preference Prediction & Personalized Recommendation based on Item Dependency Map

*Sun-Hee Youm
Digital Media R&D Center
Samsung Electronics CO., LTD

Abstract

In this' paper, we intend to find user's TV program choosing pattern and, recommend programs that he/she wants. So we suggest item dependency map which express relation between chosen program. Using an algorithm that we suggest, we can recommend an program, which a user has not saw yet but maybe is likely to interested in. Item dependency map is used as patterns for association in hopfield network so we can extract users'global program choosing pattern only using users' partial information. Hopfield network can extract global information from sub-information. Our algorithm can predict user's inclination and recommend an user necessary information.

I. 서론

디지털 방송 서비스가 시작된 이후 수많은 방송 컨텐츠들이 제공되고 있다. 특히 시청자들의 방송 선호도, 관심도, 시청 프로그램 히스토리 등의 자료를 기초로 시청자들이 원하는 시간에 다양한 정보를 자동으로 제공하는 것은 중요한 문제가 되고 있다.

본 논문에서는 시청자의 프로그램 시청 패턴을 분석하여 시청자가 어떠한 사용자 프로파일에 속하는지를 예측하고 관심이 있을 것 같은 프로그램을 추천하는 알고리즘을 제안하고자 한다. 본 논문에서는 Item Dependency Map이라는 방법을 제안하고 있다. 앞으로 이를 IDM이라 부르기로 한다. IDM 은 시청자들이 시청한 프로그램간의 상관관계를 표현한 행렬이다. 이를 홉필드 네트워크라는 신경망에 적용하여 시청자의 시청패턴을 예측하고 또한 예측된 결과를 이용하여 시청자에게 효과적이고 새로운 정보를 제공하고자 한다.

본 논문의 순서는 II에서 알고리즘에 사용할 홉필드 네트워크와 IDM에 대해서 설명하고 III, IV에서 이를 토대로 어떻게 시청자 시청 패턴을 추출하는지를 설명한다.

II. 관련 연구

2.1 홉필드 네트워크

홉필드 네트웍은 각 노드사이의 연결가중치에 기억하고자 하는 것들을 연상시킨 뒤 어떤 입력을 통해서 전체 네트워크가 어떤 평형상태에 도달하는 방식으로 작동되는 신경망 중의 하나이다. 홉필드 네트웍의 기본 구조는 입력과 연결 가중치를 곱한 값들을 모두 더해서 적당한 임계함수를 통해 출력하는 노드들이 여러 개 있고 이들이 상호 연결되어 있는 구조이다. 다른 신경망과의 차이는 출력 값이 다시 입력되는 구조로 이루어진다. 즉 일부 다른 신경망들이 단일방향으로 작동하는 정적인 반면에, 홉필드 네트웍은 시간에 따라서 내부상태가 동적으로 변화된다.

각 노드사이의 연결 가중치에 기억하고자 하는 것들을 연상시킨 후에 어떠한 입력에 대하여 출력을 반복적으로 구하면서 그 값이 더 이상 변화되지 않을 때 중단하게 된다.

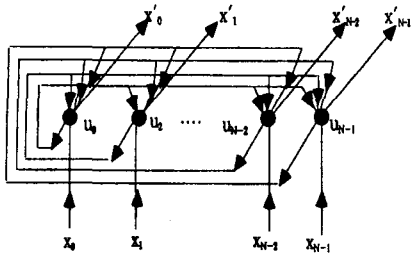

그림 1. 홉필드 네트워크의 기본구조.

2.2. IDM 알고리즘

IDM 알고리즘은 시청자들이 시청한 프로그램을 분석하여 각 시청자 유형별로 패턴이 있는 프로그램-프로그램 형태의 행렬로 만드는 것이다. IDM 알고리즘의 입력값을 만들기 위해 사전단계로서 각 시청자자들의 시청 트랜잭션을 그룹별로 나눈다. IDM이 그룹별로 생성되어야 하기 때문이다. 시청자 트랜잭션은 기본적으로 그룹 아이디, 사용자 아이디, 시청한 시간 그리고 시청한 프로그램으로 구성한다. 그룹별로 나누어진 각각의 시청자 트랜잭션은 다시 시청자별로 프로그램 시퀀스를 형성하게 된다. 프로그램 시퀀스는 각 시청자가 시청한 프로그램을 시간 순서대로 배열한 집합이다.

IDM 알고리즘은 전체 2가지 프로세스로 나뉜다. 첫 번째 프로세스는 pre_IDM을 만드는 과정으로 pre_IDM은 IDM을 만들기 전의 전처리 과정에서 생성되는 행렬이다. 프로그램 시퀀스를 기반으로 하여 각 시청자가 프로그램을 시청한 순서를 고려하여 pre_IDM을 만들게 된다. 이때 시청된 프로그램의 시간성에 가중치를 부여하기 위해 α(0 〈α≤1)를 사용하게 된다. 이 과정을 α Chain Rule이라 한다.

NI : 전체 아이템의 수

$P_k(i,j)$: k 번째 그룹의 pre_IDM

NL_kNLk : k번째 그룹의 프로그램 시퀀스 개수

$D_t(i,j)$: t 번째 프로그램 시퀀스의 프로그램 i 시청과 프로그램 j 의 시청사이의 거리

$S_k(i)$: k 번째 그룹에서 시청된 프로그램 i의 총 수

$$P_k(i,j) = \frac{\sum_{t=1}^{NL_k} \alpha^{D_t(i,j)-1}}{S_k(i)}, \quad if \ i \neq j$$

$$P_k(i,j) = 0, \quad if \ i = j$$

두 번째 프로세스는 IDM을 생성하는 과정이다. IDM을 만드는 과정을 Mapping Rule이라 한다. IDM은 각 시청자들의 시청 패턴을 표현한 행렬이다. pre_IDM을 IDM으로 변환하기 위해 본 논문에서는 두 가지 방법을 제시한다. 한 가지 방법은 행렬의 평균을 이용하는 것이고 다른 한 가지는 행렬의 각 값들간의 순위를 이용하는 것이다. 행렬과 순위는 pre_IDM의 행렬 값이 아이템간의 상관성을 나타내 줄 수 있는가 아닌가를 결정해주는 역할을 하게 된다. 마지막으로 각 그룹의 IDM이 각각 서로 다른 패턴을 가질 수 있도록 행렬을 재조정 해야 한다. 각 그룹의 행렬 형태가 일정한 패턴이 나타나지 않고 모두 전체적으로 산만하게 뿌려진 형태를 가지고 있거나 각각의 그룹들이 비슷한 형태를 나타나게 된다면 홉필드 네트워크에 학습의 효과가 나타날 수 없다. 이러한 경우 개인 사용자가 특정 그룹에 속하는 지를 찾을 수 없게 되기 때문이다.

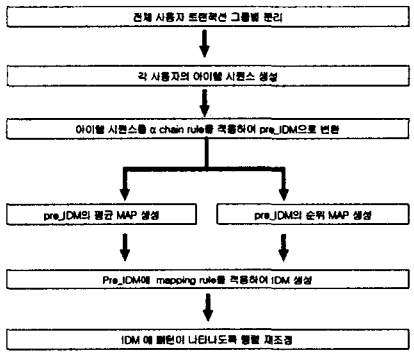

그림 3 IDM의 흐름도

III. IDM을 이용한 추천 시스템

3.1 IDM을 이용한 추천 시스템의 전체 구성

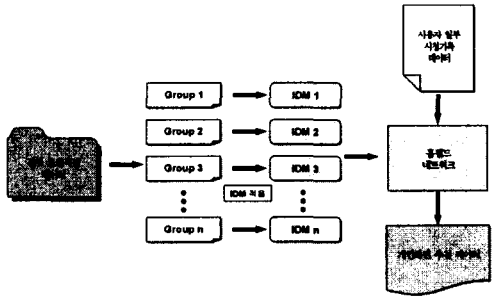

그림 4. IDM기반 추천 시스템의 전체 구성도

IDM을 홉필드 네트워크에 적용하여 사용자 구매패턴을 예측하는 추천 시스템은 그림 3과 같다. 시청자가

시청한 프로그램들은 데이터 베이스에 시청한 시간, 시청한 사용자의 ID와 같은 개인 프로파일, 시청한 프로그램의 장르 등이 저장된다. 그리고 나면 일정한 기준에 의해 시청자들은 그룹별로 분리가 된다. 시청자 그룹 아이디는 시청자의 나이, 직업, 취미 등 시청자가 직접 입력한 내용을 바탕으로 정해질 수도 있다. 또는 클러스터링(clustering)과 같은 데이터 마이닝 기법을 이용하여 시청자를 그룹별로 분리한 후에 이를 바탕으로 시청자의 그룹 아이디를 저장할 수도 있다.

IDM을 만들기 위한 기본 데이터는 프로그램 시퀀스이다. 프로그램 시퀀스는 시청자의 그룹 아이디, 시청자 아이디, 시청 시간, 시청한 프록그램으로 구성된다.

3.2 사용자 입력 데이터 변환

앞에서 설명했던 IDM알고리즘을 이용하여 각 그룹의 IDM을 만드는 과정은 다음과 같다.

표 1은 프로그램 시퀀스를 생성하기 위한 시청한 시청 데이터의 예를 보인 것이다.

그룹 1에 속한 시청자는 A와 B이고, 그룹 2에 속한 시청자는 C와 D이다. 그리고 각각의 시청된 프로그램에는 시청 시간 아이디가 있기 때문에 이를 바탕으로 시간 순서성을 가진 프로그램 시퀀스를 표 2와 표 3과 같이 만들 수 있다.

이렇게 완성된 프로그램 시퀀스가 앞장에서 제안했던 IDM 알고리즘에 의해 각 그룹별로 프로그램-프로그램의 형태인 IDM으로 변환된다. IDM은 시청 상관성이 높은 프로그램들의 원소에는 1로 그렇지 않은 프로그램들의 원소 값은 0으로 정해져 일정한 패턴을 가지게 된다. 이 패턴들은 각 그룹별로 고유한 형태를 가지게 되고, 개인 정보가 없는 사용자의 일부 시청 데이터가 입력되었을 때 예측 가능하도록 한다.

표 1. 시청자 시청 데이터의 예

시청자 아이디	시청 그룹 아이디	시청 시간 아이디	시청된 프로그램
A	1	355	1
A	1	477	22
B	1	157	44
B	1	405	12
B	1	235	16
C	2	274	39
C	2	138	20
C	2	249	11
D	2	554	23
D	2	198	41

표 2. 그룹 1의 프로그램 시퀀스

시청자 아이디	프로그램 시퀀스
1	Q_1 = (1, 22)
2	Q_2 = (44, 16, 12)

표 3. 그룹 2의 프로그램 시퀀스

시청자 아이디	프로그램 시퀀스
3	Q_3 = (20, 11, 39)
4	Q_4 = (41, 23)

그러나 각 그룹별 IDM이 고유의 형태를 가지지 않을 경우 IDM을 완성하는 단계에서 행렬 간 변환을 해 주어 각 그룹별 고유한 형태를 가지도록 해 주어야 한다.

대개 개별 사용자의 시청 데이터는 데이터베이스에 일부만이 존재할 뿐이다. 따라서 입력할 개별 시청자의 시청데이터에 α Chain Rule을 적용할 경우 시청자의 일부 시청패턴으로 나타낼 만큼의 프로그램-프로그램 행렬 형태가 나타나지 않을 수 있다. 또한 프로그램 수가 많지 않기 때문에 $D_k(i,j)$가 많이 크지 않을 것이다. 그래서 시청자 프로그램-프로그램 행렬에서는 α값을 1로 설정하여 $D_k(i,j)$가 2이상인 경우에도 원소 값을 1로 매겨 계산되게 된다. 개별 시청자 시청 데이터는 전체 IDM 알고리즘을 적용하지 않고 α가 1인 경우의 α Chain Rule 만을 적용하여 프로그램-프로그램 행렬을 만든다. 앞으로 이 행렬을 P_IDM이라고 칭하겠다.

3.2 IDM의 홉필드 네트워크 적용

IDM과 P_IDM을 학습시키기 위해서 우선 $NI \times NI$ 행렬을 $1 \times (NI \times NI)$형태의 행렬로 변환하여 학습시킨다. NI는 사용된 전체 프로그램의 수를 뜻한다. 홉필드 네트워크 학습에서 첫 번째 단계는 연결강도 W_{ij}를 결정하는 것이다. 연결강도는 아래와 같다.

$$W_{ij} = \sum_{k=1}^{N_k} V^k{}_i \; V^k{}_j$$

W_{ij}는 노드 i에서 노드 j로 연결된 연결강도이고, 그 값이 1이거나 -1인 $V^k{}_i$는 k번째 그룹에 대한 패턴

의 i번째 요소이다. 따라서 연결강도 W_{ij}는 $(NI \times NI) \times (NI \times NI)$의 정방행렬이 된다.

두 번째 단계는 P_IDM을 초기화하여 출력 값을 계산하는 과정이다. P_IDM의 i번째 노드를 결정하기 위해서는 $W_{i,j}$와 P_IDM의 각 노드 값을 곱하여 그 합이 임계값(threshold)보다 큰지 아닌지를 비교한 후에 temp값을 작으면 -1로 크지 않으면 1로 한다. 그리고 나서 P_IDM의 i번째 노드와 temp값과 비교하여 같은지 틀린지를 본다. 각 노드에 관한 이러한 계산은 모두 수행되어야 한다. 이 값이 일정회수 이상동안 더 이상 변하지 않을 때까지 각 노드의 작동은 반복된다.

Ⅳ. 사용자 프로파일 예측 및 추천

마지막으로 IDM과 P_IDM을 홉필드 네트워크에 학습시켜 나온 결과를 이용하여 시청자의 프로파일을 예측하고 시청자에게 추천을 하게 된다. 홉필드 네트워크에 P_IDM을 학습시킨 결과는 그림 4의 예와 같이 프로그램-프로그램 형태의 행렬로 표현된다. 특정 시청자의 일부 시청 데이터가 입력되어 앞으로의 시청 패턴이 예측된 행렬이다. 따라서 행렬의 원소 중 1의 값을 가지는 것은 프로그램간의 시청 상관성이 높다고 말할 수 있는 것이다. 그림4에서 ①의 경우 시청자가 프로그램 3을 시청한 경우는 프로그램 2를 시청할 상관성이 높다고 판단되므로 관리자는 이 시청자에게 추천하여 시청를 추천할 수 있을 것이다. ②의 경우는 프로그램 간의 상관성이 0으로 예측되었다. 따라서 시청자가 프로그램 3을 시청하였을 경우 프로그램 NI는 추천을 해도 시청자가 프로그램을 시청할 가능성은 적다고 판단할 수 있는 것이다. 또한 예측된 패턴과 학습된 여러 개의 IDM 중에서 비슷한 모양을 가진 IDM의 구룹이 입력된 시청자와 같은 사용자 프로파일을 가진다고 예측할 수 있게 된다.

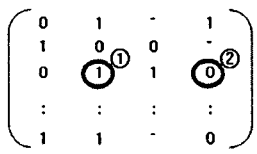

그림 4. 예측된 패턴의 예.

Ⅴ. 결론 및 향후 연구 방향

본 논문에서는 시청된 아이템간의 관계를 표현하는 IDM을 이용하여 시청자가 시청할 가능성이 있는 프로그램을 예측하여 추천하는 방법을 제시하였다. 본 논문은 추천 프로그램의 자동적인 생성이 가능하고 시청자가 시청했을 지라도 선호도 입력이 되지 않은 프로그램이 추천 목록에서 빠지게 되는 상황을 극복할 수 있는 알고리즘을 제안하였다.

그러나 생성된 각각의 IDM이 고유한 패턴을 가지게 되지 않을 경우 정확한 예측을 할 수 없는 문제점을 가지고 있다. 향후 랜덤한 방법이 아닌 방법으로 IDM을 재배열 하여 각각의 IDMdl 서로 다른 패턴을 가질 수 있는 알고리즘을 제안하고자 한다.

참고문헌

[1] Tarun Khanna, "Foundations of Neural Networks," Addison-Wesley Publishing Company, 1990.

[2] Mehmet M. Dalkilic, Edward L. Roberston, "Information dependencies," Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 245 - 253 2000.

[3] Rosa Meo, "Theory of dependence values," ACM Trans. Database Syst. Pp. 380 - 406, Sep. 2000.

[4] Sun-Hee Youm, "Personalized Recommendation based on Item Dependency Map," IEEE International Symposium Industrial Electronics, Pp. 250 -253, June, 2001.