Recommendation of (IP)TV Programs based on Collaborative Filtering using *n*-tuple Item Clustering

EunHui Kim
School of Engineering, Information Communication University
119 Munji street, Yusung-gu, Daejeon, 305-732, Korea
82-42-866-6249
ehkim@icu.ac.kr

Shinji Pyo
School of Engineering, Information Communication University
119 Munji street, Yusung-gu, Daejeon, 305-732, Korea
82-42-866-6246
vy311@icu.ac.kr

Munchurl Kim
School of engineering, Information Communication University
119 Munji street, Yusung-gu, Daejeon, 305-732, Korea
82-42-866-6246
mkim@icu.ac.kr

ABSTRACT

With the advent of multi-channel TV services, a prohibited amount of IPTV program contents becomes available to user's sides. Furthermore, the number of IPTV service providers is rapidly increasing over internets. Therefore such information overload requires large amounts of efforts for users to search and navigate the program contents that they like to watch. In this paper, we incorporate collective intelligence by utilizing collaborative filtering (CF) for grouping community users to recommend user's preferred (IP)TV programs so that the users can easily find and watch their preferred contents with significantly reduced efforts. The CF is performed in conjunction with a previous probabilistic framework but incorporates similarity-based grouping in terms of the attributes of IPTV programs for which each program is modeled as an *n*-tuple item with genre, actors/actress, channels, emission times, user's ages and genders etc. So the item-user relevance of each user group becomes trustier. We also incorporate the time weighting factor into program preference computation which leads to reflecting timely changing user preference on (IP)TV programs. Experimental results show that proposed scheme yields 86.4% of prediction accuracy for top 5 recommended programs.

Categories and Subject Descriptors

I.2.1 [**Applications and Expert Systems**]: Industrial automation – *Personalized automatic EPG processing*.

General Terms

Algorithms, Experimentation, Performance

Keywords

Collaborative filtering, IPTV, Personalization

1. INTRODUCTION

From the user's standpoint, explosive increase of (IP)TV program contents via on-air broadcasting, VOD, and online UCC bothers the users to fast find and browse their preferred contents. From content provider's sides, it is difficult to efficiently expose their contents to the target users which are difficult to be identified. Therefore it is essential to have reasoning methods that can predict and recommend user's preferred contents in personalized ways. In general, the recommendation systems automatically retrieve the contents based on user's personal preferences. And much research has been carried out to study the recommendation systems for particular problem domains.

The rating based systems utilize the user's preference based on the explicit rating values and implicitly user's purchase list [1, 3, 4, 5]. They are adequate for internet environments. In TV watching environments, log based implicit rating by user's TV watching history is more adequate [2].

It is known that in collaborative filtering (CF) more satisfied recommendation for a particular user can be achieved based on the items with ratings from similar multiple users than one single user. But CF has a drawback since a new item, which is never rated by users, cannot be predicted since there is no information about user's preference on it [3]. Also, one of the major issues of CF is computational complexity for a large number of users. Since the memory based methods in CF are to find similar preference users from database to an active user, it usually causes memory capacity problems. Xue *et. al*. incorporated a two-step clustering method with a and smoothing technique to improve clustering accuracy by dealing with rating sparseness [1]. Another way of solving computational complexity takes into account the co-occurrence purchasing items because the number of the co-occurrence preference items is much smaller than that of similar users of the whole users [3].

JunWang *et. al*. performed CF under peer-to-peer TV environments by using user-item relevance modeling [2]. Since all users' information is hardly to be known in peer-to-peer environments, the buddies in messengers are assumed as similar taste users. This method provides improved extraction of similar preference users than randomly chosen users. But, mostly IPTV environment is walled garden state with server-client structure and grouping similar preference users by measuring similarity of preference items (intrinsic interest about an item itself, genre and channel interests) gives higher correlated users, thus resulting in better recommendation. So in this paper, the server-client is considered.

This paper proposes an implicit preference reasoning method with the idea that (1) timely changing user's interest on items, (2) the item model with multiple attributes such as genre, channel, actor, actress, producer etc., and (3) user preference model on items with interests in items' attributes.

This paper is organized as follows: In Section 2, we propose a novel CF technique more suitable for IPTV program recommendation by incorporating the attributes of items (IPTV program contents); Section 3 shows experiment results and discussions; we conclude our work in Section 4.

2. Collaborative Filtering with Item Attribute

2.1 Proposed scheme

In order to make reasoning on recommendation of preferred items (IPTV program contents), it is necessary to collect the usage history of the items that were watched by all users. The usage history of the consumed (watched) items can be recorded in a database (DB) at server sides. Figure 1 shows the architecture of our proposed IPTV program recommendation system with CF.

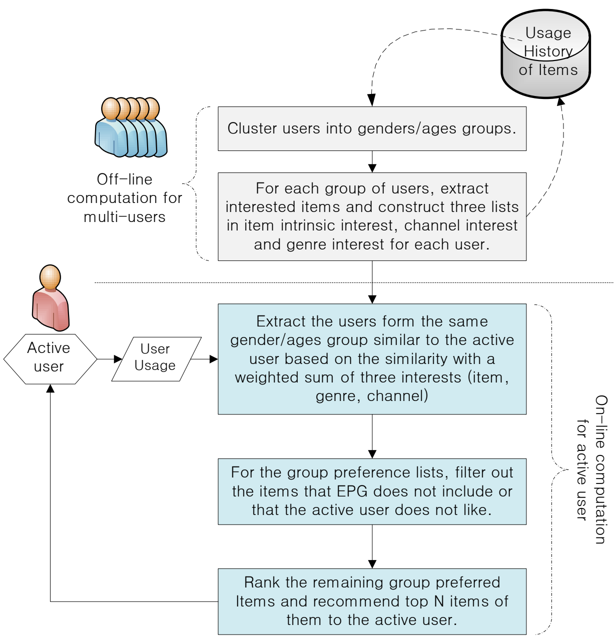

Figure 1 A flowchart of the proposed scheme

Our proposed scheme for program recommendation consists of two stages: off-line computation for multi-users and on-line computation for active users. For off-line computation, the users in DB are clustered into gender-ages groups. For each group, a set of interested items is extracted and three lists of interest are constructed in terms of item intrinsic interest, genre interest and channel interest. For on-line computation, when an active user is logged in, the corresponding similar users to the active user are extracted based on the similarity of a weighted sum of three interests. For the extracted similar users, their consumed items are compared with EPG and then the items that are also in EPG are only ranked and presented to the active user. It is noted that by comparing ranking-based item lists with EPG, precision accuracies can be improved. New items, which have never been watched and rated by users, can be also included in the recommendation lists by building user interests in item properties, such as genres and channels.

2.2 Reasoning Implicit User Preference

With the time weighting effect, we define the implicit user preference as the *user k*'s intrinsic interest in *item m* such as

$$g_k(m) = \frac{\sum_{t\in t_k} watchDuration_k(m)\cdot f(t)}{programDuration(m)\cdot freq(m)} \tag{1}$$

where $f(t)=\exp\{-|t_{now}-t|/|window|\}$, t_{now} is the current date, t is the date that the program series was watched and $freq(m)$ is the number of rerun times for a TV program (item).

In this paper, we propose to characterize each item with one or more attributes and to compute user preference on items based on the weighted sum of user's interests in attributes. So we model an *item* m (i_m) with N attributes, that is, in *n*-tuple, $i_m = (x^1(m), x^2(m), \cdots, x^N(m))$ where $x^j(m)$ is the *j-th* attribute of *item m*. On the other hand, we define $x_k^{(j)}(m)$ as the *user k*'s interest on the *j-th* attribute of *item m*. Then, the implicit user preference $y_k(m)$ on *item m* by *user k* is defined as a weight sum of the user interests on $x_k^{(j)}(m)$, and is given by

$$y_k(m) = \alpha_1 x_k^{(1)}(m) + \alpha_2 x_k^{(2)}(m) + \cdots + \alpha_N x_k^{(N)}(m) \tag{2}$$

where α_j is a weight for the user's interest on the *j-th* attribute and we have $\sum_{j=1}^{N}\alpha_j = 1$. The range of $y_k(m)$ is then between 0 and 1. In this paper, the user's interests are computed on the three attributes for each item: the interest on *item m* itself - $x_k^{(1)}(m)$, the interest in genres - $x_k^{(2)}(m)$ and the interest in channel - $x_k^{(3)}(m)$. For $x_k^{(1)}(m)$, the user's generic interest in *item m* in Eq. (1) is normalized to

$$x_k^{(1)'}(m) = \frac{g_k(m) - \bar{g}_k(m)}{\left(\underset{s'\in\{1,2,\cdots,S\}}{\arg}\max(g_k(s'))\right) - \bar{g}_k(m)} \tag{3}$$

where $\bar{g}_k(m)$ is the average of the *user k*'s intrinsic interests in all the items that exceed the threshold. So, $0 \le x_k^{(1)}(m) \le 1$. This normalization balances the range of intrinsic interest values between heavy and light TV viewers as well as the ranges of the interest values of other attributes. The genre interest $x_k^{(2)}(m)\big|_{i_m\in g_s}$ of *user k* in *item m* (i_m) belonging to genre g_s is defined as a normalized sum of the user's intrinsic interest on i_m and is given by

$$x_k^{(2)}(m)\Big|_{i_m\in g_s} = \frac{\sum_{i_{m'}\in g_s} g_k(m')}{\underset{s'\in\{1,2,\cdots,S\}}{\arg}\max\left(\sum_{i_{m'}\in g_{s'}} g_k(m')\right)} \tag{4}$$

where S is the total number of genres. Similarly, we can formulate the user's interest $x_k^{(3)}(i_m \in c_q)$ in channel c_q as the same way like genre interest.

It can be noticed that the proposed implicit user preference on *item m* can be used not only to remove the items with their implicit user preference values less than a predefined threshold but also to rank the items in order. The items that their item intrinsic interest values $x_k^i(m)$ exceed the average (T_k) are extracted into the recommendation candidate list. The average of the item intrinsic values is computed as

$$T_k = \frac{1}{M}\sum_{m'=1}^{M} x_k^1(m') \tag{5}$$

where M is the total number of the items that the *user k* has watched. By taking the items that their implicit user preference values exceed the average value, similarity-based clustering results becomes more reliable than simply using a small number for T. Therefore the items that their implicit user preference values exceed the average value are called preferred items by users.

2.3 User Grouping by Similar Preferences

Grouping users based on gender-ages is useful to analyze the consumption behaviors of users on TV program contents in different ages-genders [1]. In Section 2.1 we first cluster the users into different gender-ages groups. And then, similar users to an active user are searched in the same gender-ages group based on PCC metric. We compute the PCC values by taking into account the attributes of each item.

The PCC between an active user (u_a) and each user (u) in the same gender-ages group is given by

$$sim_{u_a,u} = \sum_{j=1}^{N}\beta_j \frac{\sum_{i_m \in L_{u_a}} \left(x_u^{(j)}(m) - \bar{x}_u^{(j)}(m)\right)\cdot\left(x_{u_a}^{(j)}(m) - \bar{x}_{u_a}^{(j)}(m)\right)}{\sqrt{\sum_{i_m \in L_{u_a}}\left(x_u^{(j)}(m) - \bar{x}_u^{(j)}(m)\right)^2}\sqrt{\sum_{i_m \in L_{u_a}}\left(x_{u_a}^{(j)}(m) - \bar{x}_{u_a}^{(j)}(m)\right)^2}} \tag{6}$$

where N is the total number of attributes of an item. In Eq. (8), $\beta_j\left(\sum_{j=1}^{N}\beta_j = 1\right)$ is a weight for a user u's interest $x_u^{(j)}(m)$ on the *j-th* attribute of *item m*. u_a is an active user and L_{u_a} represents a list of the active user's preferred items (programs). And $\bar{x}_u^{(j)}$ is the average value of *user u*'s interest in the *j-th* attribute of an item.

The range of $sim_{u_a,u}$ value is $[-1,1]$. So, the items for which $sim_{u_a,u} > 0$ are extracted for a candidate list of item recommendation.

2.4 Construction of A Candidate Recommendation List

After similar users to active user are found in Section 2.3, we construct three candidate recommendation lists in terms of intrinsic item interest, genre interest and channel interest for the active user u_k with the union of the items that the similar users prefer to. Since the candidate recommendation list of intrinsic item interest may include some programs which are no longer broadcast, those programs are excluded by comparing with EPG at the time instance of recommendation.

EPG may include new items (programs) that have not been consumed (watched) by the active user and similar preference users. In this case, those new items would never be included in the recommendation list. In order to overcome this drawback, the lists of genre interests, $x_k^{(2)}(m)\big|_{i_m \in g_s}$ and channel interests $x_k^{(3)}(i_m \in c_q)$ of each user is computed and saved. Therefore, there are some chances that the new items in EPG can be included in the candidate recommended list. In general, the numbers of users who show similar interests in item's genres and channels are very large. To deal with all items by such users becomes computationally prohibitive. Therefore, in this experiment, we only take 20~50% of all similar users for recommending new items.

2.5 Ranking of Relevant Items for Recommendation

Jun Wang *et. el*. modeled the relevance between user and item based on language modeling for text retrieval method [3], which will be our basis for recommending the relevant items (programs) to users. The item based generation model of relevance rank is

$$RANK_{u_a}(i_m) = \left(\sum_{\substack{i_b \in L_{u_a} \cap \\ c(i_b, i_m) > 0}} \log\left(1 + \frac{(1-\lambda)P(i_b \mid i_m, r)}{\lambda P(i_b \mid r)}\right)\right) + \log P(i_m \mid r) \tag{7}$$

where L_{u_a} indicates the profile list of user u_a, i_b is an item belonging to L_{u_a} and $c(i_b, i_m)$ is the number of the users that showed their interests in both items i_b and i_m. The conditional probability of i_b given i_m for relevance is given by

$$P(i_b \mid i_m, r) = \frac{C(i_m, i_b)}{C(i_m)} \tag{8}$$

The first term's denominator part $P(i_b \mid r)$ in Eq. (7) is the background model to avoid the zero co-occurrence between i_b and i_m, i.e., $P(i_b \mid i_m, r)$ due to sparseness of user-item matrix.

3. Experiments and Results

For the experiment, we use a Nelson Korea data set which was collected from December 1 2002 to May 31 2003 with usage history of TV programs by 1,995 users and with the contents composed of 1,855 TV programs via 6 channels.

After we assume that the watching date is April 30th 04', the usage (log) data of TV programs for the first 5 months is used for training and the remaining data of one month is used for test.

The performance is measured by precision. Precision is the ratio that how many relevant items exist in the returned item. In this paper, the relevant items are selected the same as choosing preference items. In other words, for test duration, only the items which interest value exceeds average interest value are chosen as relevant items.

3.1 Precision results for λ

For different λ values, it can be seen in Figure 2 that the precision performance is not affected by λ except large values ($\lambda = 1$). Since similar users for each active user are selected as a group, the group preference becomes quite similar to that of the active user. Therefore λ doesn't affect much.

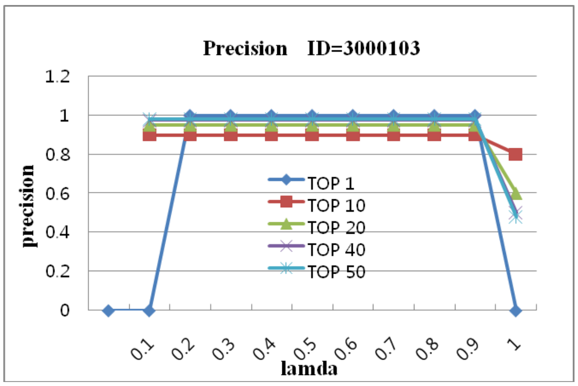

Figure 2 Precision results for different λ values

For $\lambda = 1$, the precision performance becomes very poor because the group's preferred items are only considered without active user's preferred items for recommendation. In the case of $\lambda < 1$, the precision performance turns out to be good because the combination of maximum likelihood model $P(i_b \mid i_m, r)$ and background model $P(i_b \mid r)$ works properly.

3.2 Precisions performance with α and β

By the experiments, we found that best performance of the proposed scheme is achieved with the interest weights in order such that the weight for the interest in the first attribute (item itself) has the largest values against the other two weights such as $(\alpha_1 = 0.8, \alpha_2 = 0.15, \alpha_3 = 0.05)$. This order of weight values seems to reasonably reflect the user's preference with the most specific attribute (item itself) to the least specific attribute (channel) for different user's distinction. For the weight of the PCC measures in Eq. (6), the performance is best achieved with $(\beta_1 = 0, \beta_2 = 0, \beta_3 = 1)$. This is because, based on the channel interest, the proposed scheme is able to collect more amount of users so that there are more chances for the relevant items to be recommended for the active users. And it also decreases computation complexity since channel dimension is smaller than item dimensions.

For a set of the fixed weights $(\alpha_1 = 0.8, \alpha_2 = 0.15, \alpha_3 = 0.05)$ and $(\beta_1 = 0, \beta_2 = 0, \beta_3 = 1)$, the precision results of the recommended items by the proposed method are compared with those of a traditional method for randomly chosen 20 active users from each gender-ages group shown in Table 1. Notice that the traditional method is to simply cluster the similar user groups based only upon the items. The proposed recommendation method yields higher precision performance than the traditional one. It is interesting to see that the proposed method exhibits prominent outperformance against the traditional method for the old-age (50'~60's) female groups which show more consistent consumption of TV programs (items) than other age groups.

Table 1 Precision comparison for each user with top *N* items

Profile Group	Recommendation Methods	Numbers of recommended items			
		TOP 1	TOP 5	TOP 10	TOP 20
Female 60's	Traditional	1	0.9	0.79	0.805
	Proposed	1.000	0.920	0.875	0.878
Female 50's	Traditional	1.000	0.970	0.840	0.755
	Proposed	1.000	0.970	0.915	0.843
Female 30's	Traditional	0.813	0.750	0.638	0.578
	Proposed	0.938	0.763	0.694	0.614
Male 20's	Traditional	0.933	0.787	0.573	0.601
	Proposed	0.933	0.787	0.813	0.717
Female 10's	Traditional	0.950	0.770	0.710	0.615
	Proposed	1.000	0.880	0.835	0.840
Average	Traditional	0.939	0.835	0.710	0.671
	Proposed	0.974	0.864	0.826	0.778

4. Conclusion

In this paper, we propose a CF technique with a probabilistic framework by incorporates into the similarity-based grouping the attributes of TV programs such as genre, actors/actress, channels, emission times, user's ages and genders. For this, user preference on TV programs (items) is modeled with multiple attributes so that the similar user grouping for CF becomes more relevant to active users. Experimental results show that our proposed scheme is computationally efficient and yields more than 86.4% of prediction accuracy for top 5 recommended programs.

REFERENCES

[1] Xue, G., Lin, C., Yang, Q., Xi, W., Zeng, H., Yu, Y., Chen, Z. 2005. Scalable collaborative filtering using cluster-based smoothing. In Proceedings of ACM SIGIR, 114~121.

[2] Wang, J., Powelse, J., Fokker, J., Vreies, A., Reinders, M. 2007. Personalization on a peer-to-peer television system. Multimedia Tools and Applications. vol. 36, no. 1-2 (Jan. 2007), 89-113.

[3] Deshpande, M., Karypis, G. 2004. Item based Top-N recommendation algorithms. ACM Transactions on Information Systems, vol. 22, issue 1 (Jan. 2004). 143-177.

[4] Ding, Y., Li, Xue. 2005. Time Weight Collaborative Filtering. In Proceedings of the 14th ACM international conference on Information and knowledge management (Oct. 2005). 485 – 492.

[5] Breese, J. S., Heckerman, D. Kadie, C. 1998. Empirical analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 43~52.