# Cold-Start Group Profiling with a Clustering-Coupled Topic Model

Zhijian Jiang[1], Yanfeng Wang[1](✉), Weiyuan Chen[1], Xie Wang[2], Ya Zhang[1],
Jianping Mei[3], and Zhuowei Huang[3]

[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University,
Shanghai, China
wangyanfeng@sjtu.edu.cn

[2] National Engineering Research Center of Digital Television, Shanghai, China

[3] China Central Television, Beijing, China

**Abstract.** While interactive television enables a new user-centered TV mode, catering to the tastes of TV users is one of the most critical tasks in delivering interactive TV experience. It faces two key challenges. First, the user behaviors on TV are much sparser than those of the internet users, thus making the modeling of user preferences more challenging. Second, an TV account is usually associated with multiple individuals in a family, making it difficult to discriminate the preferences of individual family members. In this paper, we thus propose a novel Clustering-Coupled Topic Model (CCTM), which characterizes user profile only by analyzing user viewing behaviors without any program metadata. This model clusters the users into different groups, then access the group preference for program recommendation by coupling the interest of different users in the same cluster group. We validate the performance of the CCTM with real-world data from a national interactive TV program. The experimental results have demonstrated that the CCTM can reasonably extract the users' potential preference, which is further leveraged to recommend programs to the users.

**Keywords:** User profile · Group recommendation
Interactive television · User view behaviours
Clustering-Coupled Topic Model

## 1 Introduction

Digital interactive television provides more interactive experience in terms of applications, services and interactions, compared to traditional television [1]. It also raises the opportunity of delivering personalized content to the viewer [2]. Thus, television suppliers need to identify model user behavior by analyzing user log [3–5]. However on user profile there are 2 key challenges - data sparsity and multi-user mix problem. The inconvenient operations on television remote control limit user's viewing resulting in the Sparse user behavior. On the contrary,

G. Zhai et al. (Eds.): IFTC 2017, CCIS 815, pp. 334–342, 2018.
https://doi.org/10.1007/978-981-10-8108-8_31

the convenience of Internet page jumps is accompanied by a lot of clicks. Multi-user mix problem means that an TV account is usually shared by more than one members in a family who may have different preferences, which make it difficult to model each member. In practice, user behavior logs are usually treated as documents, in which user is analogous to document and each viewing content is analogous to words in corresponding documents. Thus, topic model [6] can be applied to the documents to model the user preferences and content in vectors. In addition, document clustering [7] can also be utilized to organize similar users into groups, which is significant for user modeling and user preference visualization. Basically, we can perform user clustering method, such as K-means and spectral clustering [8], based on user vectors generated by topic model. Applying them separately fails to capture the coherence between topic and cluster. Recent work [9] shows that applying topic model for document clustering significantly improves the performance of topic model.

The assumption is reliable that correlation the programs a user frequently watches reveal the his/her preference. The analysis for interactive television users' behaviour will be helpful for catering to the tastes of TV users. However, there are some difficulties brought by the sparse user log and the constitutional complexity of an TV account. In this paper, we propose a novel Clustering-Coupled Topic Model (CCTM) in order to model interactive television users' behaviour. Our chief contribution are as follows: (1) We develop a novel Clustering-Coupled Topic Model which incorporate topic model and document clustering. And we derive approximate inference for the model. (2) Based on the proposed model, we extract the user preference from user log and then recommend programs to users.

We experiment the proposed model with a real-world data set provided by a national interactive TV in China. Our analysis has revealed that proposed CCTM is able to cluster users with coherent interest into groups. There is a promising improvement in evaluation index when we further leverage the proposed CCTM for personalized program recommendation.

The rest of the paper is organized as follows. Section 2 introduce the LDA model and the Clustering-Coupled Topic Model. Section 3 introduce the real data and the experiments analysis. Finally, Sect. 4 conclude the paper.

## 2 Clustering-Coupled Topic Model for Group Profiling

In this section we will introduce the Clustering-Coupled Topic Model for Interactive television user behaviors, which leverages the viewing behaviors to give user profile. Firstly, We introduce how LDA and kmeans may be used to cluster users as well as group profiling. Then we present the structure of the proposed clustering-coupled topic model, including the generative process of the model and the variational inference method.

### 2.1 LDA Model and Cluster Algorithm Help Group Profiling

A hypothesis is proposed that user viewing process is the embodiment of user interest, an interactive account responding to a family with various interests, and each interest may be expressed as the distribution of television programs. Based on such hypothesis, the LDA model will be adapted to fit the process of user viewing which can be divided into two-step. First, a single interest is selected based on an interest distribution that is sampled from a Dirichlet distribution. Second, a TV program is chosen according to the multinomial distribution that the chosen interest over programs. As result, the generative process for each viewing behaviour is presented below:

(1) Choose $\theta_d \sim Dir(\alpha)$, where the $Dir(\alpha)$ is the Dirichlet distribution of the parameter $\alpha$.
(2) For each viewing $n$, where $\forall 1 \leq n \leq N_d$.
   (a) choose an interest $z_{d,n} \sim Multinomial(\theta_d)$.
   (b) choose a TV program $w_{d,n}$ from $p(w_{d,n}|z_{d,n}, \beta)$, a multinomial conditional probability where the interest is $z_{d,n}$.

The Fig. 1 shows the details of the graphical model of LDA. With the LDA model, each family account will be represented by a vector by the distribution over the interests, where each interest is responsible for generating the TV programs.

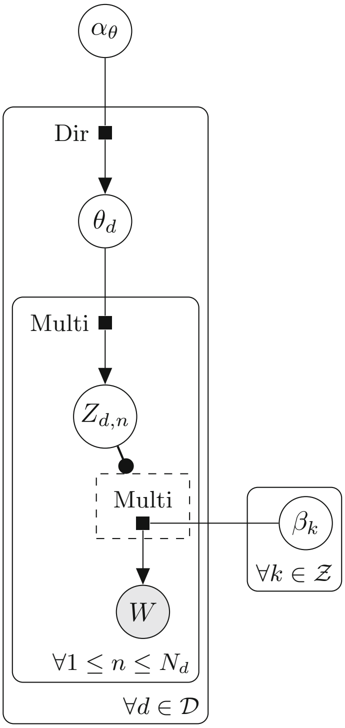
$\alpha_\theta$
Dir
$\theta_d$
Multi
$Z_{d,n}$
Multi
$\beta_k$
$\forall k \in \mathcal{Z}$
$W$
$\forall 1 \leq n \leq N_d$
$\forall d \in \mathcal{D}$

**Fig. 1.** Latent Dirichlet Allocation (LDA) topic model

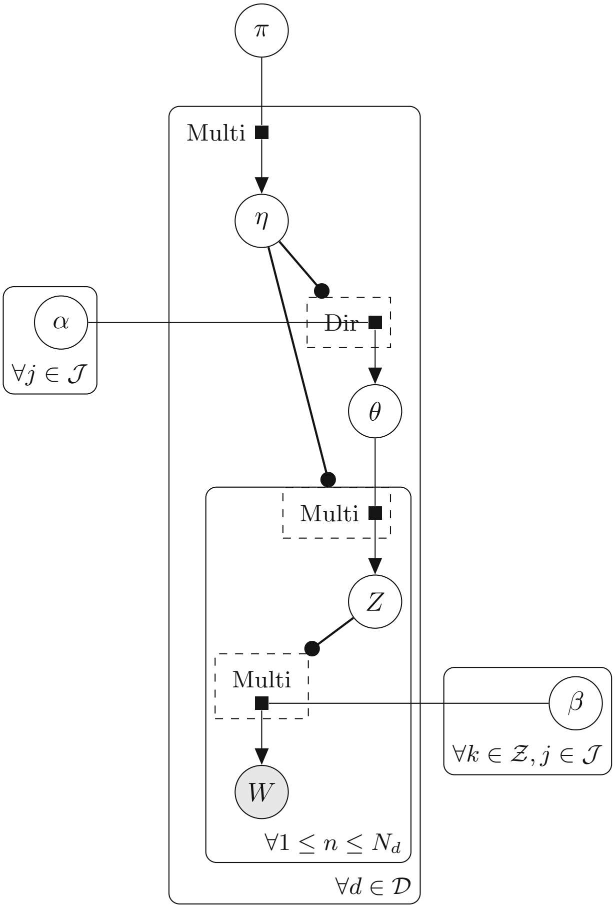
$\pi$
Multi
$\eta$
$\alpha$
$\forall j \in \mathcal{J}$
Dir
$\theta$
Multi
$Z$
Multi
$\beta$
$\forall k \in \mathcal{Z}, j \in \mathcal{J}$
$W$
$\forall 1 \leq n \leq N_d$
$\forall d \in \mathcal{D}$

**Fig. 2.** Clustering-coupled topic model

The results of LDA model facilitate clustering, the K-means algorithm can be using in the interest space to obtain clusters. As result, the group profiling is characterized under the help of LDA model and K-means. However, this strategy that performing LDA model and cluster separately ignores their interaction and the chicken-and-egg relationship. Performing the topic models and clustering jointly contributes to promote each other to realize the global optimal.

### 2.2 Clustering-Coupled Topic Model

The clustering-coupled topic model is shown in Fig. 2. Given a corpus containing $\mathcal{D}$ user accounts $d \in \mathcal{D}$, there is a assume that user accounts inherently belong to $\mathcal{J}$ groups $j \in \mathcal{J}$. Each group $j$ possess a group-specific Dirichlet prior parameter $\alpha_j$ as well as $K$ group-specific local topics $\beta_{j,k}$.

With the clustering-group LDA model, the viewing behaviors of a family account are generated through a three-step process. First, for each family, associated with a group indicator, a group $\eta_d$ is sampled from the multinomial distribution parametrized by $\pi$. Second, we sample a local proportion vector $\theta_{\eta_d}$ from the group Dirichlet prior $\alpha_{\eta_d}$ on the condition where group $j = \eta_d$. Third, for each viewing $n$, an interest $z_{d,n,\eta_d}$ is picked up according to the local proportion vector $\theta_{\eta_d}$, and then a program $w_{d,n}$ is generated from the a multinomial conditional probability $p(w_{d,n}|z_{d,n,\eta_d}, \beta_{j,k})$. The generative process of CCTM can be summarized as follows:

(1) For each account, sample a group $\eta_d \sim Multi(\pi)$.
(2) Sample group interest proportion $\theta_{\eta_d} \sim Dir(\alpha_{\eta_d})$.
(3) For each viewing behaviour.
  (a) sample a group interest $z_{d,n,\eta_d} \sim Multinomial(\theta_{\eta_d})$.
  (b) sample a TV program $w_{d,n} \sim Multinomial(\beta_{j,k})$

### 2.3 Variational Inference and Parameter Learning

Using variational inference [10], we perform approximate inference help solve the critical inference problem what estimating the posterior distribution $p(H|w, \Pi)$,where $H = \{\eta, \theta, z\}$ refer to latent variables, $w$ are observed variables and model parameters are $\Pi = \{\pi, \alpha, \beta\}$. The basic idea of variational inference is emply another distribution $q(H|\Omega)$ to approximate the true posterior $p(H|w, \Pi)$ by minimizing their Kullback-Leibler (KL) divergence, where model parameters $\Omega = \{\zeta, \mu, \phi\}$. The model of distribution $q(H|\Omega)$ is shown in Fig. 3, the mathematical formulation is

$$\begin{aligned} & q(H|\Omega) \\ & = q(\eta, \theta, z|\zeta, \mu, \phi) \\ & = q(\eta|\zeta) \textstyle\prod_{j=1}^{J} q(\theta_j|\mu_j) \prod_{i=1}^{N} \prod_{j=1}^{J} q(z_{i,j}|\phi_{i,j}) \end{aligned} \tag{1}$$

where $\zeta$, $\phi_{i,j}$ are multinomial parameters, $\mu_j$ is Dirichlet parameter.

The process of variational inference is divided into E-step and M-step, and the result are shown as follows.

In E-step, we optimize the variational parameters while keeping model parameters fixed

$$\begin{aligned} \zeta_j \propto \pi_j exp\{log\Gamma(\textstyle\sum_{k=1}^{K} \alpha_{j,k}) - \sum_{k=1}^{K} log(\Gamma\alpha_{j,k}) \\ + \textstyle\sum_{k=1}^{K} (\alpha_{j,k} - 1)[\Psi(\mu_{j,k}) - \Psi(\sum_{t=1}^{K} \mu_{j,t})]\} \end{aligned} \tag{2}$$

$$\mu_{j,k} = \zeta_j(\alpha_{j,k} - 1) + \textstyle\sum_{n=1}^{N} \zeta_j \phi_{j,i,k} + 1 \tag{3}$$

$$\phi_{j,i,k} \propto exp\{\Psi(\mu_{j,k}) - \Psi(\textstyle\sum_{t=1}^{K} \mu_{j,t}) + \sum_{v=1}^{V} w_i^v log\beta_{j,k}^v\} \tag{4}$$

In M-step, the variational parameters is fixed, we optimize the model parameters by maximizing the lower bound

$$\pi_j \propto \textstyle\sum_{d=1}^{D} \zeta_{d,j} \tag{5}$$

$$\beta_{j,k,v} \propto \textstyle\sum_{d=1}^{D} \sum_{i=1}^{N_d} \zeta_j \phi_{d,i,j,k} w_{d,i}^v \tag{6}$$

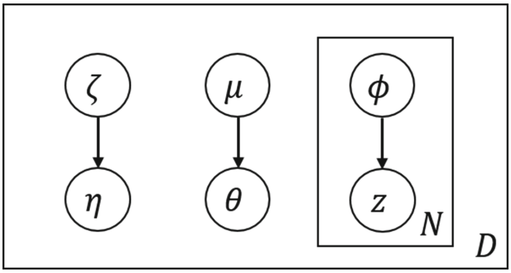

**Fig. 3.** Approximate model of clustering-coupled topic model

## 3 Datasets and Experiments

We validate our model on a real-world data set provided by a national interactive TV in China. The entire data covers the user from January 2017 to July of the viewing behaviour, and contains 200,025 families and 9,580 TV programs.

The Fig. 4 and Table 1 show the user viewing curve, less than 20% users posses more than 6 viewings and more than half of the user's viewing behavior is less than twice. Then the Fig. 5 and Table 2 describe the times that the program is viewed, more than half of programs posses less than 50 clicks. These tables and curves clearly illustrate that most people view few program and most programs possess lower click times. In fact, the user's behaviour in this data set is much sparser than the Movielens, with the sparsity of 0.000618, which is a serious problem to analyze user preference.

The Clustering-Coupled Topic Model is proposed here to model the sparse user log. Table 3 shows some actual interactive activities. Each record contains the family account ID (CAID), the program ID (PID), the program title (TITLE) and the program channel (CHANNEL), while only user account and program ID are needed in the Clustering-Coupled Topic Model.

In experiments, the number of groups $J$ is set to 20 and the number of topics every group $K$ is set to 10. Meanwhile other parameters are randomly initialized. The CCTM clusters the users into different group: Fig. 6 shows some group profiling by the form of presenting the most associated programs. Obviously, different groups have different preferences for different types of programs. For example, the group on the right side of the picture consists of the programs that most are provided from channel CCTV8. However, the group users represented on the upper-lift corner are more likely to be related to the programs from CCTV1 channel.

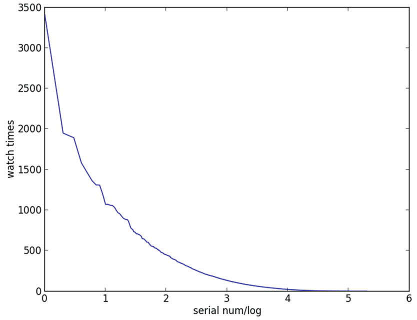

**Fig. 4.** The curve about the user viewing counts

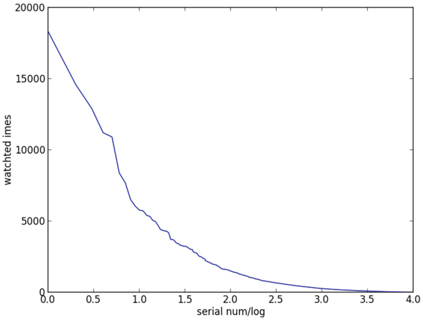

**Fig. 5.** The curve about the viewed times of every program

**Table 1.** User average viewing counts

| Segment ratio | View counts |
|---|---|
| 0% - 10% | 43.11 |
| 10% - 20% | 6.91 |
| 20% - 30% | 3.72 |
| 30% - 40% | 2.51 |
| 40% - 50% | 2.01 |
| 50% - 60% | 1.26 |
| 60% - 70% | 1.23 |
| 70% - 80% | 1.12 |
| 80% - 90% | 1.03 |
| 90% -100% | 1.01 |

**Table 2.** Watched times of program

| Segment ratio | Watched times |
|---|---|
| 0% - 10% | 830.4 |
| 10% - 20% | 189.8 |
| 20% - 30% | 112.2 |
| 30% - 40% | 73.7 |
| 40% - 50% | 50.8 |
| 50% - 60% | 33.1 |
| 60% - 70% | 19.3 |
| 70% - 80% | 10.1 |
| 80% - 90% | 4.1 |
| 90% -100% | 1.4 |

**Table 3.** Example records of interactive activities

| | CAID | PID | TITLE | CHANNEL |
|---|---|---|---|---|
| 1 | Ox12324321348953 | VIDA1468554199296462 | Journey to the West | CCTV8 |
| 2 | Ox12893223432134 | VIDA1468567043464536 | Tale of White Snake | CCTV11 |
| 3 | Ox13424238957845 | VIDE1468566577362380 | Snooker | CCTV5 |

**Table 4.** Recommendation classical evaluation index

| | Precision | Recall | F-Measure |
|---|---|---|---|
| CCTM | 0.046308 | 0.093905 | 0.047008 |
| LDA | 0.027296 | 0.062598 | 0.028389 |
| Hot | 0.002301 | 0.005536 | 0.002387 |
| Random | 0.000791 | 0.001570 | 0.000747 |

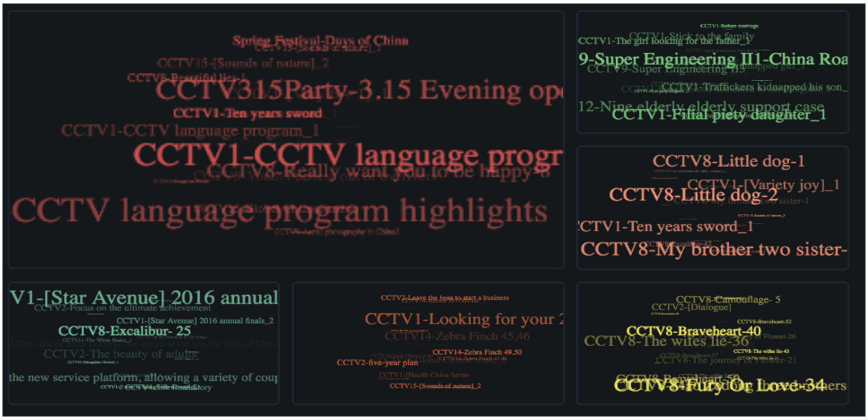

**Fig. 6.** Group profiling represented by program title

Furthermore, we apply the result of group profiling for program recommendation. The Table 4 shows that the CCTM has a good improvement on the classical evaluation index including precision, recall or F-Measure when compared to the baseline. Random means the random program recommendation, and the Hot is defined that providing the most popular programs for everyone. LDA refers to the personalized recommendation according to the classical topic model [6]. Apparently, the CCTM achieves the best performance in program recommendation.

## 4 Conclusion

We propose a novel Clustering-Coupled Topic Model, which takes into account various user interactive viewing behaviours to access implicit user preferences. The largest advantage of this model is the compatibility that only user behaviours are required. Even with the lack of any program metadata or family structure, this model would still obtain user interests from their viewing, and then cluster different users into different group and characterize group profiling by the vector of interests or programs. Furthermore, we validate the performance of the CCTM with real-world data from a national interactive TV. The experimental results have demonstrated that the CCTM can reasonably extract the users' potential preference, which can be further leveraged to recommend programs to the users.

In the future, one direction of this study is extend our model to semi-supervised or supervised clustering settings, for the reason the prior knowledge on family structure is sometimes available.

## References

1. Jan, D.V., Peters, O., Heuvelman, A.: Interactive television or enhanced television? The Dutch users interest in applications of ITV via set-top boxes. In: Annual Meeting of the International Communication Association ICA (2017)
2. Cho, J.H., Sah, Y.J., Ryu, J.: A new content-related advertising model for interactive television. In: IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, pp. 1–9. IEEE (2008)
3. Branch, P., Egan, G., Tonkin, B.: Modeling interactive behavior of a video based multimedia system. In: Proceedings of the IEEE International Conference on Communications, pp. 978–982 (1999)
4. Gopalakrishnan, V., Jana, R., Knag, R., Ramakrishnan, K., Swayne, D., Vaishampayan, V.: Characterizing interactive behavior in a large-scale operational IPTV environment. In: Proceedings IEEE INFOCOM 2010, pp. 1–5 (2010)
5. Zhang, Y., Chen, W., Zha, H., et al.: A time-topic coupled LDA model for IPTV user behaviors. IEEE Trans. Broadcasting **61**(1), 56–65 (2015)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 9931022 (2003)
7. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_4
8. Ng, A.Y., Jordan, M.I., Weiss, Y., et al.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 2, pp. 849–856 (2002)
9. Xie, P., Xing, E.P.: Integrating document clustering and topic modeling. In: Proceedings of the 29th International Conference on Uncertainty in Artificial Intelligence (2013)
10. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends@ Mach. Learn. **1**(1–2), 1–305 (2008)