Mining IPTV User Behaviors with a Coupled LDA Model

Weiyuan Chen, Ya Zhang, and Hongyuan Zha

***Abstract*—In this paper, we analyze user behaviors in Internet Protocol television (IPTV). We propose a novel coupled LDA model, which considers topic of TV programs viewed as well as the timestamps of the viewing behaviors, in order to capture the inherent viewing patterns of the users along the topic as well as the time dimensions. Based on the coupled LDA model, we further summarize the behavior patterns of IPTV users. We compare our model with other models on the real data collected from an operational nation-wide IPTV system and show that the proposed coupled LDA model is able to capture interesting viewing patterns.**

***Index Terms*—IPTV, User Behavior Analysis, Coupled LDA**

I. Introduction

INTERNET Protocol television(IPTV), the television services through the Internet, has recently gained increasing popularity. The interactivity inherent to IPTV makes it possible for service providers to observe user activities. As a result, many efforts were made to understanding the behaviors of IPTV users. The user viewing activity was investigated for VOD based systems [1][2]. User's channel switching activities were studied to improve the efficiency of IPTV systems in [3][4][5]. Other researchers studied recommender systems to help users filter items of interest, including recommending preferred TV programs or contextual advertisement [6][7][8].

Modeling user activities in an IPTV network is essential for delivering personalized experiences such as program recommendation and personalized contextual advertisement, a differentiating feature of IPTV services over traditional TV services. But so far little work has been done on user behavior modeling. In this paper, we propose a model to characterize the behaviors of IPTV users. We focus on the usage of VOD functions, i.e. each family selectively plays their preferred TV programs at a certain time. Each record contains the family ID (FID), the starting time of a program, the program ID (PID) and the program title. Here different episodes of the same TV shows have different program ID but same title. The example records are shown in Table I.

The programs viewed by a family in general manifest the family's interests. The following procedure is assumed for program selection. First, the interests of a family are sampled from a Dirichlet distribution. Second, for each viewing activity, a single interest is chosen from the interest distribution of the family. Third, a TV program is selected from a multinomial distribution over TV programs specific to the chosen interest. It is hence straightforward to apply Latent Dirichlet allocation (LDA)[9] to model the interests of family.

TABLE I
IPTV EXAMPLE DATA

FID	Starting Time	PID	Program Title
196843d1bb	2011-12-30 14:35:51	440929	The Black Fox
196843d1bb	2011-12-30 15:37:41	442425	The Black Fox
1968470219	2011-12-30 20:55:10	444352	Golden Code

On the other hand, the time of the day that one watches TV could be an important indicator of an individual's demographic information. For example, it is usually impossible for people who have a daily job to watch TV during working hours. In this paper, we propose a coupled LDA model which considers topic of TV programs viewed as well as the timestamps of the viewing behaviors, in order to capture the inherent viewing patterns of the users along the topic as well as the time dimensions. Based on the proposed model, we mine behavior patterns of IPTV users and further use the mined patterns to characterize and understand family structures.

II. Proposed Model

In the records of IPTV user activities, in addition to TV program information, each program choice made by a family is associated with a timestamp. The IPTV data have the following characteristics:

1) A family has one or more members;
2) Each member's interests are drawn from a mixture of interests;
3) Each member tends to watch TV in certain time periods of the week, showing temporal viewing patterns.

We propose a coupled LDA (cLDA) model based on the above characteristics. We extend the interest distribution $\vec{\theta}_m$ in standard LDA model to the interest over time distribution. Table II shows a toy example of distribution $\vec{\theta}_m$ in LDA model and cLDA. In LDA, each family has exactly one interest distribution during all time. While in cLDA, each family has different interest distributions in different time spans. With cLDA model, we automatically group similar TV programs into an interest, and similar timestamps into a time span (temporal

Weiyuan Chen and Ya Zhang are with the Institute of Image Communication and Network Engineering & Shanghai Key Laboratory of Multimedia Processing and Transmissions, Shanghai Jiao Tong University (e-mail: {cwyalpha, ya_zhang}@sjtu.edu.cn). Ya Zhang is the corresponding author.

Hongyuan Zha is with the School of Computational Science and Engineering, Georgia Institute of Technology (e-mail: zha@cc.gatech.edu).

pattern). Here the interests over time are defined as the activity patterns.

TABLE II
INTEREST DISTRIBUTION $\vec{\theta}_m$ OF FAMILY m

(A) $\vec{\theta}_m$ in LDA

Interest	Distribution
Cartoon	0.3
Variety	0.4
War	0.3

(B) $\vec{\theta}_m$ in cLDA

	17~19PM weekdays	20~22PM weekdays	13~16PM weekdays	13~18PM weekends
Cartoon	0.18	0.01	0.01	0.1
Variety	0.01	0.2	0.15	0.04
War	0.01	0.1	0.01	0.18

According to the cLDA model, the viewing behaviors of a family are generated through a four step process. First, for each family, a distribution of activity pattern is sampled from a Dirichlet distribution. Second, for each viewing activity of the family, an interest and a temporal pattern are chosen from the distribution. Third, a TV program is selected from a multinomial distribution over TV programs specific to the chosen interest. Fourth, a timestamp is selected from a multinomial distribution over timestamps specific to the chosen temporal pattern.

TABLE III
NOTATION USED IN THIS PAPER

Symbol	Description	Type
K	number of interests	Scalar
L	number of temporal patterns	Scalar
M	number of families	Scalar
V_w	number of unique TV programs	Scalar
V_t	number of unique timestamp	Scalar
N_m	number of behaviors in family m	Scalar
$\vec{\theta}_m$	the multinomial distribution of activity patterns for the family m . $\Theta = \{\vec{\theta}_m\}_{m=1}^{M}$	KL dimensional vector
$\vec{\phi}_k$	the multinomial distribution of TV programs to the interest k	V_w dimensional vector
$\vec{\psi}_l$	the multinomial distribution of timestamps to the temporal pattern l	V_t dimensional vector
$z_{m,n}$	the activity pattern associated with the n -th activity in the family m	$[1, \mathrm{KL}]$
$w_{m,n}$	the TV program associated with the n -th activity in the family m	$[1, V_w]$
$t_{m,n}$	the timestamp associated with the n -th activity in the family m	$[1, V_t]$
α, β, γ	Dirichlet priors	Scalar

Each family is characterized by a multinomial distribution of activity pattern. Using the notation in Table III, we reshape the activity pattern distribution matrix of size $K \times L$ to a KL dimensional vector $\vec{\theta}_m$. Thus the z th component of the vector $\vec{\theta}_m$ represents probability of activity that the family selects interest $z_{m,n,2}$ in temporal pattern $z_{m,n,1}$, where

$$z_{m,n,1} = ((z_{m,n} - 1) / K + 1), z_{m,n,2} = (z_{m,n} \mod K) \quad (1)$$

For each family m, its program selection process using the cLDA model can be shown as follows:

1) Choose $\theta_m \sim Dir(\alpha)$ where $m \in \{1, \cdots, M\}$ and $Dir(\alpha)$ is the Dirichlet distribution for parameter α
2) choose $\phi_k \sim Dir(\beta)$ where $k \in \{1, \cdots, K\}$
3) choose $\psi_l \sim Dir(\gamma)$ where $l \in \{1, \cdots, L\}$
4) For each activity n, where $n \in \{1, \cdots, N_m\}$
   a) choose $z_{m,n} \sim Multinomial(\theta_m)$
   b) choose a timestamp $t_{m,n} \sim Multinomial(\psi_{z_{m,n,1}})$
   c) choose a TV show $w_{m,n} \sim Multinomial(\phi_{z_{m,n,2}})$

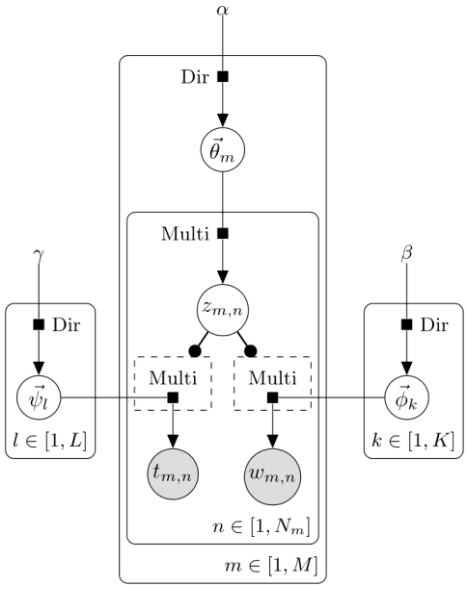

Fig. 1. The Graphical Model of cLDA.

The graphical model of cLDA is shown in Fig 1.

We perform approximate inference for cLDA using Gibbs sampling. In a single sampling iteration, we sample an assignment $z_{m,n}$ for an individual activity ($w_{m,n}$, $t_{m,n}$), conditioned on other assignments $\vec{z}_{\neg(m,n)}$ for all activities except for ($w_{m,n}$, $t_{m,n}$). The conditional distribution is as follow:

$$P(z_{m,n} \mid \vec{w}, t, \vec{z}_{\neg(m,n)}, \alpha, \beta, \gamma) \propto \frac{n_{z_{m,n,2}}^{(w_{m,n})} + \beta - 1}{\sum_{v_w=1}^{V_w} (n_{z_{m,n,2}}^{(v_w)} + \beta) - 1}$$

$$\times \frac{n_{z_{m,n,1}}^{(t_{m,n})} + \gamma - 1}{\sum_{v_t=1}^{V_t} (n_{z_{m,n,1}}^{(v_t)} + \gamma) - 1} \times (n_m^{(z_{m,n})} + \alpha - 1) \quad (2)$$

where $n_{z_{m,n,2}}^{(v_w)}$ is the times of TV program v_w assigned to interest $z_{m,n,2}$. $n_{z_{m,n,1}}^{(v_t)}$ is the times of timestamp v_l assigned to temporal pattern $z_{m,n,1}$, $n_m^{(z_{m,n})}$ is the number of activities in family m assigned to the activity pattern $z_{m,n}$.

III. IPTV DATASET

We validate the proposed cLDA model with a real-world data set provided by an operational nation-wide IPTV system in China. It contains 288,497 families and 1,830 TV programs, where different episodes of the same TV program are considered as one single program. Here we restricted the IPTV dataset to users that have at least 20 activities. The data set contains 34,974,827 records by 154,622 families on 1,805 TV programs from April 1st to December 31th 2011. The view count statistics for families and TV programs are shown in Figure 2 and Figure 3. The point (x, y) in Figure 2 means that there are y families that have exactly x activities. The point (x, y) in Figure 3 means that there are y TV programs that have been viewed at least x times.

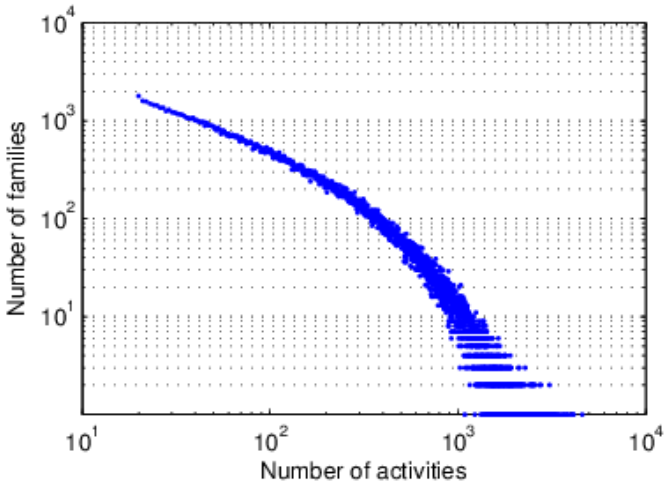

Fig. 2. The Long Tail of Family Activity

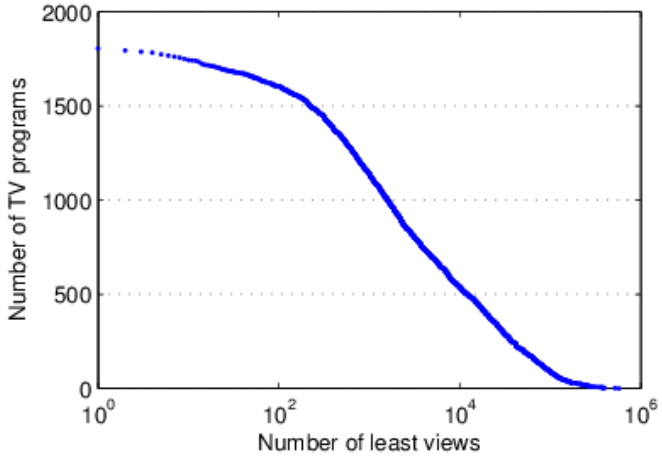

Fig. 3. The Long Tail of TV Program Popularity

IV. EXPERIMENTS

Here we represent the timestamp in the form of `weekday_hour'. In the following experiments, we compare the proposed cLDA model with LDA and the Topics over Time (TOT) model [10]. We empirically set the parameters as follows: $K = 50, L = 8$. For TOT model, we change the Beta distribution to multinomial distribution to fit the characteristics of the IPTV data.

A. Interest Discovered

When examining the TV programs associated with an interest group generated by LDA, we find that extremely popular TV programs tend to dominate an interest, resulting in that other TV programs in the same interest group are irrelevant to the popular one. Table IV shows such an interest group generated by LDA

TABLE IV
AN INTEREST IN LDA

Program Title	Probability
Palace: The Locked Heart Jade	0.7848
Opposite Attraction III	0.0439
Opposite Attraction II	0.0409

TABLE V
AN INTEREST IN CLDA

Program Title	Probability
Palace: The Locked Heart Jade	0.5728
Schemes of a Beauty	0.3627
Happy Mother-in-law, Pretty Daughter-in-law	0.0492

which contains top three TV programs and their generating probability from the interest group. The popular show `Palace: The Locked Heart Jade' with a high probability dominates the interest. However, `Palace' is a Chinese romance historical fiction show, and `Opposite Attraction' is a Korean modern love show. There is little coherence in the interest group.

Table V shows the interest group containing the popular show `Palace' in cLDA. Compared to the interest group generated by LDA, the set of TV programs generated by cLDA are all Chinese romance historical fiction shows. Their generating probability tends to be more even.

Those interest groups dominated by a few TV programs are found to be less coherent. Thus we define the low-coherence interest group as the one dominated by less than 10 TV programs. More precisely, for each interest group we first arrange the TV programs in the descending order of their generating probability and then select the top 10 programs from them. If there are two consecutive TV programs that the generating probability of the previous one is over 10 times bigger than the following one, we say this interest is dominated by less than 10 TV programs. LDA produces 9 low-coherence interests, while cLDA only has 5, which shows that cLDA can produce more coherence interests by utilizing the temporal information.

TABLE VII
PATTERNS DISTRIBUTION GENERATED BY CLDA

Temporal Pattern	0	1	2	3	4	5	6	7
Cartoon for infant	0.0006	0.0006	**0.3042**	0.0006	0.0006	0.0006	0.0006	**0.0530**
Cartoon for child	0.0006	0.0006	**0.2343**	0.0006	0.0006	0.0025	0.0006	**0.0388**
Chinese War Show	**0.0271**	0.0090	0.0084	0.0006	0.0161	0.0006	0.0006	0.0006

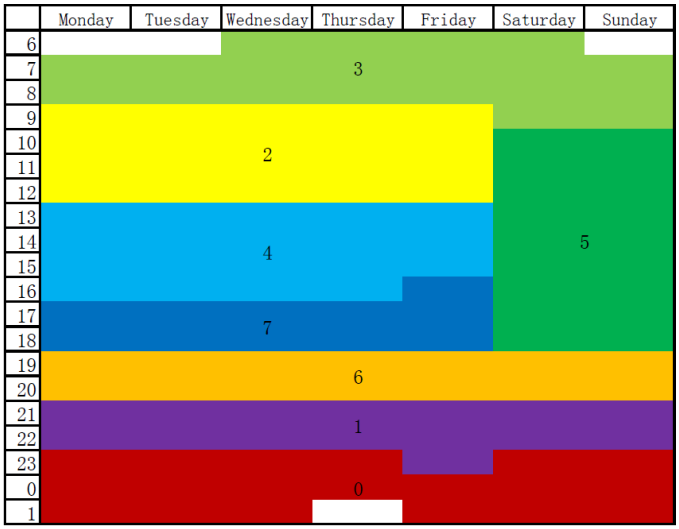

Fig. 4. The Temporal Pattern of cLDA

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
6	0	0	0	0	0	1	0
7	5	7	1	0	1	0	0
8	14	25	26	10	20	3	7
9	12	9	29	9	23	5	2
10	13	26	33	21	29	4	1
11	28	33	36	39	47	8	3
12	23	24	27	33	27	6	4
13	1	2	4	1	0	0	1
14	0	0	0	0	1	0	2
15	10	12	9	14	7	1	3
16	19	32	7	18	20	5	4
17	9	7	4	7	4	5	14
18	8	6	1	5	4	12	9
19	3	0	2	0	3	6	0
20	1	0	2	0	1	0	2

Fig. 5. Cartoon's Watching Frequency of A Family

B. *Temporal Pattern Discovered*

The temporal pattern generated by cLDA is shown in Fig 4, where hours belong to the same pattern is in the same color. The temporal pattern clearly separates the time span of a weekday into morning, afternoon, evening and midnight.

Pattern 3 (light green) suggests that there is a temporal viewing pattern at early morning of every day. It is interesting to see that the pattern include 9am from weekends to this "early morning" pattern but not 9am from weekdays, suggesting cLDA model is able to capture the fact that people may get up later and leave home later in weekends and weekdays.

Pattern 7 (blue) most likely corresponds to the temporal pattern of students during weekdays. It starts one hour early at Friday. A possible explanation is that class in generally ends earlier at Fridays.

The daytime of the weekends is treated as a single temporal pattern (Pattern 5, green). This captures the viewing pattern of some people who tends to watch TV all day during weekends.

Another interesting pattern is Pattern 1 (purple). This pattern captures the behaviors of those who watch TV during 9pm and 10pm. It is interesting to see that for Fridays, 11pm is also included in the pattern. A possible explanation is that individuals tends to stay up later to watch TV.

C. *Case Study*

All the three methods, LDA, TOT, and cLDA, target to predict the kind of TV programs a family would like to watch. But TOT and cLDA are able to further predict the type of TV shows a family would like to watch at a certain time. The difference between TOT model and cLDA model is that the TOT generates the pattern only if most families have such watching behaviors.

Here we use a family that mainly watches cartoon in the weekday morning as an example to demonstrate the difference. Figure 5 shows the cartoon programs' watching frequency of the family. The integers in the Table represent the frequency of the family watching cartoon in a particular time. For example, the family watches cartoon programs 47 times at 11AM of Friday. The data clearly show that the family mainly watches cartoon in the weekday morning.

The patterns of the family generated by TOT are shown in Table VI. TOT can only assign the global watching patterns for the particular family. As can be seen from the table, TOT fails to tell the cartoon's watching time for the family.

TABLE VI
PATTERNS DISTRIBUTION GENERATED BY TOT

Interest	Temporal Pattern	Probability
Cartoon for infant	5~7PM everyday	0.3527
Cartoon for child	5~7PM everyday	0.2737
Chinese War Show	8~11PM everyday	0.0796

The patterns produced by TOT include both interests and temporal patterns. TOT can produce the patterns only if many families watches the same kind of TV in a similar time span. In our dataset, children in most family tend to watch cartoon after school at 5 ~ 7PM. Thus the TOT can only generate the global patterns of watching cartoons in 5~7PM.

Table VII shows the pattern distributions generated by cLDA, where the temporal pattern 0~7 can be found in Fig 4. For example, the temporal pattern 2 in the Table means 9~12AM in weekdays. cLDA successfully tells that the family likes to watch

TABLE VIII
PREDICTIVE-PERPLEXITY

K	5	10	20	30	40	50
LDA	428.5	360.2	297.1	254.7	239.7	225.3
TOT	415.8	367.7	296.1	267.6	239.6	231.8
cLDA	388.0	352.6	268.2	248.8	224.9	206.5

cartoons at 9~12AM in weekdays. cLDA can produce more customized pattern for each family.

D. Perplexity Analysis

We also perform collaborative filtering task to predict what the TV show the family would choose when they turn on their TV at a certain time. The predictive-perplexity on M_{test} test families is defined as below:

$$predictive-perplexity(D_{test}) = \exp\left\{-\frac{\sum_{m=1}^{M_{test}} \log p(w_{m,N_m} \mid \vec{w}_{m,1:N_m-1}, t_{m,1:N_m})}{M_{test}}\right\} \quad (3)$$

A lower perplexity score indicates a better generalization performance. In these experiments, the parameter K for each model ranges from 5 to 50, and the parameter L for the cLDA is 6. Table VIII shows that cLDA performs best on the predictive-perplexity.

E. Family Classification

In this section, we use K-means clusters to group the families into 10 clusters by their distribution $\vec{\theta}_m$ obtained from three models.

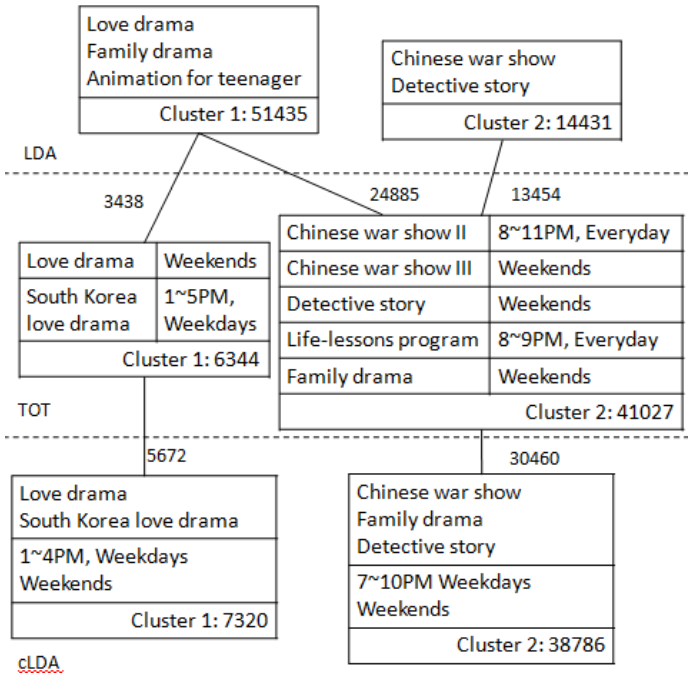

Fig. 6. Clustering Families on Love Drama and War

Figure 6 shows example clustering results for each model. In each cluster table, the top interest and the number of the cluster is presented. The number attached to the line between clusters generated by different model indicates the number of families that appear both in the two clusters. For example, cluster 2 generated by LDA has 13,454 families in common with cluster 2 generated by TOT. Moreover, the top timestamps associated with the pattern in TOT are shown in the table. When clustering the distribution $\vec{\theta}_m$ for cLDA, we find that interests in each cluster have a similar temporal pattern distribution. Thus in the table of cLDA, we only present the most common temporal pattern for the cluster.

In Figure 6, the cluster 1 of TOT and cluster 1 of cLDA contain almost the same families. These families tend to watch love drama in the afternoon at weekdays. They are probably housewife who needn't work in the day time. Without the temporal information, LDA mixes this kind of families in its cluster 1. cLDA and TOT may help us find out a new kind of family role.

V. Conclusion

IPTV has brought users more interactive and personalized TV viewing experiences. Understanding user preferences is the critical in delivering personalized IPTV services. We propose in this paper a novel coupled LDA model, which considers interests of TV programs viewed as well as the timestamps of the viewing behaviors, in order to capture the inherent viewing patterns of the users along the interest as well as the time dimensions. Based on the coupled LDA model, we further summarize the behavior patterns of IPTV users. We perform an in-depth study on several intrinsic characteristics of IPTV user activities by analyzing the real data collected from an operational nation-wide IPTV system. Our analysis has revealed that the coupled LDA model is able to reveal temporal patterns in TV watching behaviors as well as coherent interest groups.

Acknowledgment

This work was supported in part by the High Technology Research and Development Program of China (2011AA01A107, 2012AA011702)， Shanghai Science and Technology Rising Star Program (11QA1403500), National Natural Science Foundation of China (61221001), and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

References

[1] P. Branch, G. Egan, and B. Tonkin. Modeling interactive behavior of a video based multimedia system. In Proceedings of the IEEE International Conference on Communications, pages 978-982, 1999.

[2] V. Gopalakrishnan, R. Jana, R. Knag, K. Ramakrishnan, D. Swayne, and V. Vaishampayan. Characterizing interactive behavior in a large-scale operational iptv environment. In Proceedings IEEE INFOCOM 2010, pages 1-5, 2010.

[3] A. Abdollahpouri, B. Wolfinger, J. Lai, and C. Vinti. Elaboration and formal description of iptv user models and their application to iptv system analysis. In MMBnet2011, 2011.

[4] T. Adomkus, R. Bruzgiene, and L. Narbutaite. Influence of users behaviour to iptv service. Electronics and Electrical Engineering, 18(8):105–108, 2012.

[5] T. Qiu, Z. Ge, S. Lee, J. Wang, J. Xu, and Q. Zhao. Modeling user activities in a large iptv system. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, pages 430–441, 2009.

[6] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, , and C.-H. Lee. Aimed: a personalized tv recommendation system. In Proceedings of the 5th European conference on Interactive TV, 2007.

[7] J. Kim, E. Kwon, Y. Cho, and S. Kang. Recommendation system of iptv tv program using ontology and k-means clustering. Communications in Computer and Information Science, 2011.

[8] R. Konow, W. Tan, L. Loyola, J. Pereira, and N. Baloian. Recommender system for contextual advertising in iptv scenarios. In The 14th International Conference on Computer Supported Cooperative Work in Design (CSCWD2010), pages 617–622, 2010.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, Mar. 2003.

[10] X. Wang and A. McCallum. Topics over time: a non-markov continuous time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 06, pages 424–433. ACM, 2006.