Time Dependency in TV Viewer Clustering

Mengxi Xu[1,2], Shlomo Berkovsky[2], Irena Koprinska[1],
Sebastien Ardon[2], Kalina Yacef[1]

[1] University of Sydney
School of Information Technologies
NSW 2006 Australia
mexu6980@uni.sydney.edu.au
{irena.koprinska, kalina.yacef}@sydney.edu.au

[2] National ICT of Australia (NICTA)
Locked Bag 9013
Alexandria NSW 1435 Australia
{shlomo.berkovsky,sebastien.ardon}@nicta.com.au

Abstract. Web-based catch-up TV services allow users to watch programs at their favoured time and device and are revolutionizing the existing TV watching habits. With the increasing offer and demand for catch-up TV, it has become evident that there is a need for personalised recommendations that will help users to pick programs of interest from a large collection of available content. In order to mitigate the cold start problem, a catch-up TV recommender needs to exploit information pertaining to the watching patterns and stereotypical behaviour of users. This paper presents an exploratory study into the watching patterns and stability of the identified stereotypical user behavior using a large-scale dataset gathered by an Australian catch-up TV services provider. Using clustering, we were able to identify eight distinct and meaningful behaviour stereotypes. We further analysed these clusters and found that clusters with highly dominant watching patterns stabilise sooner and can be identified more accurately than others. Our work provides a solid foundation for developing future catch-up TV recommender systems.

1 Introduction

In the era of Internet Protocol TV (IPTV) and Web-based TV services, the concepts of *video-on-demand* and *catch-up TV* are taking a large and steadily increasing part in new watching practices. These services allow users to choose the programs (movies, shows, news) they prefer to watch from a vast collection of available TV content, at any time, and on a wide variety of platforms and devices. Studies indicate that users largely tend to watch stored TV programs rather than live broadcast [1] and tend to combine the two modes of delivery [11].

As with every content made available through the Web, this plethora of TV content available to users brings a major drawback: the information overload that users face when they want to select a program to watch. This problem is particularly acute in the entertainment sector, where users want to lay back and relax, and not to spend time searching for the right content [3]. This has raised a crucial need for personalised solutions and recommender systems capable of selecting a small number of relevant programs on behalf of users. Indeed, there is a lot of research activity, such as in the area of electronic program guides [8], to help users navigate through hundreds of channels. We are focusing here on a slightly different TV recommendation problem posed by the catch-up TV services.

Producing relevant catch-up TV recommendations brings a number of challenges. Firstly, the relevance of programs varies over time: since the content is available for a short period of time (typically, several weeks), new programs and programs that are about to expire could have their relevance increasing. There is also a decay effect for some types of programs, e.g., news and sports, but not for other evergreen programs, e.g., old movies and documentaries. Secondly, the user information is unreliable: in addition to catch-up TV users are likely watch free-to-air TV (not only this information is not captured, but there is also no point to recommend already watched programs) and there might be multiple users, e.g., a family, accessing one catch-up TV service through a single computer/TV, which encumbers the user modelling task. Thirdly, TV and media content often has incomplete metadata that could aggravate the application of content-based recommendation methods. Finally, it is of utmost importance to establish rapidly the trust of new users through providing recommendations that are relevant and at the same time serendipitous, in order to engage users and entice them to further use the service and consume more content.

In this work we address the cold start problem and aim to develop stereotypical user modelling and recommendations approaches, in order to mitigate the lack of sufficient user data and facilitate accurate personalisation. As the first step in this direction, we present our exploratory study that investigates the ways to group users with respect to their watching habits and analyses the evolution of these groups over time (i.e., how early they can be accurately identified), so that they can be used for stereotypical personalisation. We conducted a clustering analysis of TV logs gathered by a leading Australian national TV network, which were captured over the course of 10 weeks. The clustering identified eight distinct clusters of users. Some of them were identified and remained stable from as early as the first week of the logs, whilst others took up to nine weeks to evolve. We correlated the stability of the clusters with the dominance of certain watching patterns of users and discovered that clusters with highly dominant patterns stabilise sooner and can be indetified more accurately than others. This analysis, which we detail in the paper, is extremely valuable as a foundation for developing the catch-up TV recommender system.

The paper is structured as follows. Section 2 discusses the related work in personalisation and recommendations in the TV domain. Section 3 describes our dataset and clustering methodology that was applied. Section 4 presents and discusses our experimental results, namely which clusters were identified and how they stabilize over time. Finally, we conclude the paper and present our future research directions.

2 Related Work

A number of works explored the use of personalisation in the TV domain. The work of O'Sullivan *et al*. showed that intelligent TV personalisation is generally of a strong demand and business value [7]. The *PTVPlus* recommender system they designed involved association-rule mining and case-based methods, and proved to outperform traditional collaborative filtering recommendation methods [9]. The *Fischlar* system also showed that implicit user modelling in the TV domain is as accurate as the explicit one, which can be used to decrease the load of users [6]. Zimmerman *et al*. designed the *Touch and Drag* system that enriched TV recommendations with a usable interface [14]. They conducted a user study, which showed that the recommender engine containing both explicit and implicit learners delivers accurate personalisation to users and the interface is effective and easy to use.

Research into how to cater for the needs of a group of TV users was carried out by Masthoff [5]. The results from several experiments conducted in this work showed that users valued the fairness factor and intended to avoid personal emotions when discussing TV programs to watch. An aggregation of individual user profiles yielded the highest user satisfaction, as users' preferences towards programs changed over time. Another approach for merging multiple user profiles for the purposes of generating group-based TV recommendations was proposed and evaluated by Yu *et al*. [13]. The profile merger was based on total distance minimization function, such that the merged profile reliably reflected the preferences of the group members and the generated recommendations demonstrated high degree of classification accuracy.

The work of Ardissono *et al*. investigated the application of hybrid user modelling in the TV domain [1]. They have combined explicit and stereotypical user models to characterise user preferences towards programs and recommendations. The evaluation showed that the enrichment of user models based on community preferences, stereotypical preferences, and channel content analysis allowed to achieve a better performance than traditional user modelling. More recent work by Bellekens *et al*. introduced the *iFanzy* system supporting advanced user modelling techniques and TV recommendations functionality [2]. This work leveraged Semantic Web technologies in order to extract useful information from online social networks, which allowed to resolve the user model cold start problem. The evaluation results were shown to improve the accuracy of the models of new users.

Although numerous works have addressed the TV personalisation challenge, to the best of our knowledge only Bonnefoy *et al*. evaluated the application of clustering methods to enhance the recommendation task [4]. In that work, TV programs were clustered according to their similarity and simple interface components allowed users to indicate preferred or undesired TV content. Then, the clusters of programs were used to enrich the recommendation lists and include items similar to the preferred ones or remove the undesired items and filter out similar items. However, to the best of our knowledge no prior work involved characterisation and clustering of users based on their watching patterns. Also, the specificity of catch-up IPTV services considered in our work poses new constraints, which differ from those posed by the traditional live and on-demand TV services. This is still an open research area that raises several challenging research questions.

3 User Clustering

3.1 Dataset

The data we used in this work had been gathered by a major Australian TV broadcaster, which runs several national free-to-air TV channels. This channel offers both in-house produced and international programs, which makes it a popular service at the national level. In addition to the live broadcast, a catch-up TV service is also available through an enticing Web portal, which allows users to watch on-demand any show they may have missed. Most videos in the catch-up TV catalogue are available for a short period of time (typically, for two weeks), with some original and in-house produced programs remaining available for a longer time. The catch-up TV portal currently does not provide any personalised services to users, but the front page of the portal is curated through editorial decisions of domain experts. The curated content is segregated into three categories: featured, recently added, and will retire soon. In addition, users are able to discover TV content using traditional Web navigation paradigms: genres/categories, search engine, lists of related programs, and so on.

We captured usage logs of the Web portal for a period of ten weeks, ranging from 7-Feb-2012 to 17-Apr-2012. At a high-level, we captured more than 7 million views of 928,879 unique users, who watched altogether 3950 unique programs. Thus, every user watched on average 7.7 programs and every program was watched on average by 1812.3 users. Due to privacy limitations, very little information was available about the users: we had only access to their IP address and browser cookie number. As the IP addresses change frequently, the cookie number was considered to be the unique user identifier. Note that cookies do not necessarily uniquely identify users, but rather a Web browser (there could be several users using the same browser to access the catch-up portal and the cookies may be cleaned by users from time to time. This is an inherent limitation of the captured dataset.

The information about the TV programs included the title of the program, publication and expiry dates, season and episode information for shows, and the genre/category of the program. Thirteen program categories were set a priori by domain experts: arts, children (aged 6 to 15), comedy, documentaries, drama, education, lifestyle, news, panel, preschool (children aged under 6), reruns (children programs broadcast a long time ago), shop, and sport. Every program was classified by domain experts to a single category, which is another limitation of the dataset. Every captured view included the identity of the program that was watched, the identity of the user who watched it, and the date of the view. Due to the limitations of the logging mechanism deployed by the TV channel, no information related to the portion of the program that had been watched by users was available.

Figure 1 plots the number of views and unique users captured for every day of the logs. As can be seen, the number of views is mildly increasing and hovering around the 100,000 daily views mark, whereas the number of unique users is reasonably steady and close to 40,000. Both of them demonstrate regular peaks of watching activity, which correspond to the increased amounts of TV watching observed over the weekends. These are highlighted in the figure with dotted lines.

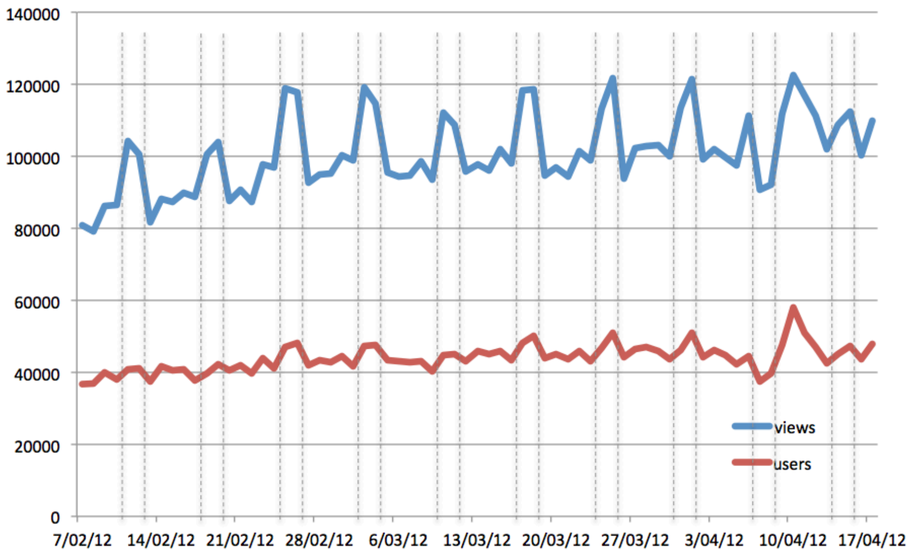

Figure 1: Number of views and users.

3.2 User Representation and Clustering

We represented every user as a 13-dimensional feature vector that captures the user preferences based on the categories of programs they previously watched. The dimensions of the vectors correspond to the above mentioned thirteen categories of TV programs and the scores of the dimensions reflect the relative number of programs of the corresponding type watched by the user. This was computed as the number of watched programs of the relevant category divided by the total number of programs watched by the user.

To identify groups of users with similar stereotypical watching patterns, we applied clustering. Each user can be considered as a point in the 13-dimensional space, which facilitates the use of distance-based clustering algorithms. We applied the well-studied K-means clustering algorithm using the Euclidean distance metric. K-means is a popular, effective and relatively efficient clustering algorithm [10, 12]. The algorithm receives the target number of clusters K as a parameter and initially selects K random points as the centroids of the clusters. Then it iteratively assigns each point to the cluster of the closest centroid (the distance is quantified using a pre-defined metric, in this case, the Euclidian distance) and re-computes the new centroids as a weighted average of the points that belong to the cluster. The process of assigning points to clusters and re-computing the centroids is repeated until the stopping criterion, e.g., no change of the centroids, is satisfied.

4 Experimental Results

In this section we present the analysis of the clusters that were identified. Initially, we will present the identified clusters and discuss the watching patterns of users in the clusters, and then we will analyse the stability of clusters over time. For the clustering, we selected a set of 110,341 users who watched 10 programs or more. Altogether, this dataset included 6,492,766 views. That is, every user in the evaluation watched on average 58.8 programs.

4.1 Identified Clusters

The first question refers to the identification of the most appropriate number of clusters, *K*. We exhaustively clustered the user profiles with the values of *K* varying from *K=5* to *K=13*, and for each *K* assessed the formed clusters using both quantitative measures (unsupervised cohesion and separation, as well as supervised entropy) and qualitative analysis [10]. The latter included domain dependent analysis of the clusters in terms of compatibility of the prevailing categories. Following these analyses, *K=8* was selected as the most appropriate number of clusters[1].

Table 1 summarises the identified clusters. Each cluster is characterised by its centroid in the 13-dimensional space. For each cluster, we list up to four prevailing categories of the centroid, limiting ourselves to categories with score greater than *0.1* only. We also show the size of the clusters, i.e., the number of users who were mapped to the cluster and the week after which the cluster evolved (will be elaborately discussed in the next sub-section).

Table 1: The identified clusters.

num	1st category		2nd category		3rd category		4th category		size	week
	category	score	category	score	category	score	category	score		
c1	drama	0.734							27,740	1
c2	docu	0.489	lifestyle	0.128	comedy	0.105			8,175	4
c3	lifestyle	0.500	docu	0.113	drama	0.107	comedy	0.106	8,462	2
c4	children	0.514	drama	0.158	preschool	0.137			9,332	3
c5	preschool	0.868							16,755	1
c6	panel	0.253	drama	0.169	comedy	0.155	lifestyle	0.108	14,698	9
c7	comedy	0.540	drama	0.139					11,485	1
c8	children	0.920							13,514	1

We briefly discuss the stereotypical watching patterns identified in these clusters. Cluster c1 is dominated by dramas and this is the cluster of drama lovers. Documentaries are prevailing in c2, followed by lifestyle programs and comedies. Lifestyle programs are prevailing in c3, followed by documentaries, dramas, and comedies. Children programs are prevailing in c4, followed by dramas and preschool programs. Cluster c5 is strongly dominated by preschool age programs and this is clearly the

[1] Detailed results of cohesion, separation, and entropy obtained for various values of *K* are omitted from the paper.

cluster of younger children. In c6, we observe a mix of panels, dramas, comedies (although all with relatively similar low scores), followed by lifestyle programs. Comedies are prevailing in c7, followed by dramas and this is the cluster of comedy lovers. Finally, cluster c8 is strongly dominated by children programs and this is clearly the cluster of older children.

Overall, we observed a balanced distribution of users across clusters with average cluster size of *13,792* users. Also the categories of programs watched in different clusters were distinct, such that the clusters provided a meaningful grouping of users with respect to their stereotypical watching patterns. Finally, these eight clusters were stable and consistently identified by the *K*-means clustering algorithm also for *K>8*. Hence, we will use *K=8* and the identified clusters as the ground truth for the following cluster stability analysis.

4.2 Stability of Clusters

As we discussed earlier, stereotypical recommendation is one the ways to mitigate the cold start problem. We will use the identified eight clusters as the basis for the user stereotypes. However, at the very initial bootstrapping stages of the system the available information may be insufficient to accurately clusters users, leading to the cold start problem of the stereotypical recommender. Hence, we will analyse the evolution of clusters and the stability of their identification over time.

For this, we repeated the *K*-means clustering procedure using the data available at the end of each week and compared the identified clusters with the clusters that were identified when all the logs were available (considered as the ground truth). Two metrics were used to assess the accuracy of clustering [12]:

$$precision(c_i, t) = \frac{|U(c_i, t) \cap U(c_i, t_f)|}{|U(c_i, t_f)|}$$

$$accuracy(t) = \frac{\sum_i |U(c_i, t) \cap U(c_i, t_f)|}{\sum_i |U(c_i, t)|}$$

In these equations, $U(c_i,t)$ denotes the set of users who were mapped to cluster c_i at the end of week t, $|U(c_i,t)|$ denotes the size of this set, and t_f denotes the timing of the complete logs, i.e., the end of week *10*. Note that $precision(c_i,t)$ is a cluster-based metric, whereas *accuracy(t)* reflects the overall accuracy of clustering across all the identified clusters. Table 2 shows the overall accuracy scores. As can be seen, the accuracy steadily improved over time. This can be explained by the increasing amount of data that was available every week, which made the clustering more reliable and more accurate.

Table 2: Average accuracy of clustering.

t	1	2	3	4	5	6	7	8	9	10
accuracy(t)	*0.546*	*0.632*	*0.669*	*0.679*	*0.683*	*0.690*	*0.699*	*0.712*	*0.960*	*1.000*

However, the observed accuracy may vary across the clusters, due to the very different nature of the underlying watching patterns. To address this question, we considered the week at which the clusters evolved, i.e., the week at which the watching patterns of the centroid became similar[2] to those found in the complete logs at the end of week *10*. It can be seen that some clusters, e.g., c2 or c6, evolved later than others (see Table 1). This is explained by the degree of dominance of certain categories within the clusters. For example, clusters c1, c5, and c8, each having only one category that scored higher than *0.1* (in fact, it scored higher than *0.734*), evolved already after one week. On the contrary, c6 had four categories that scored higher than *0.1*, but the top category had a score of *0.253* only, and the cluster evolved after nine weeks. Clusters c3, c2, and c4 had either three or four categories scoring higher than *0.1*, with the top category scoring closely to *0.5*, and they evolved after two, three, and four weeks, respectively.

In order to ascertain this dependence, we computed the correlation between the standard deviation of the set of cluster category scores (as the indicator of the uniformity of category scores) and the week the cluster evolved and stabilised. The results showed a negative correlation of *-0.756*. That is, the period of time needed for clusters with low standard deviation (and uniform category scores) to evolve was longer than the one needed for clusters with high standard deviation (and a few dominant categories) to evolve.

We were also interested to analyse the time-based fluctuations in the precision scores achieved by the 8 clusters. Figure 2 plots the individual precision scores of the clusters. The horizontal axis represents the 10 weeks and the vertical – the precision scores for that week. It should be highlighted that we plot the precision curves of a cluster starting only from the week that the cluster evolved.

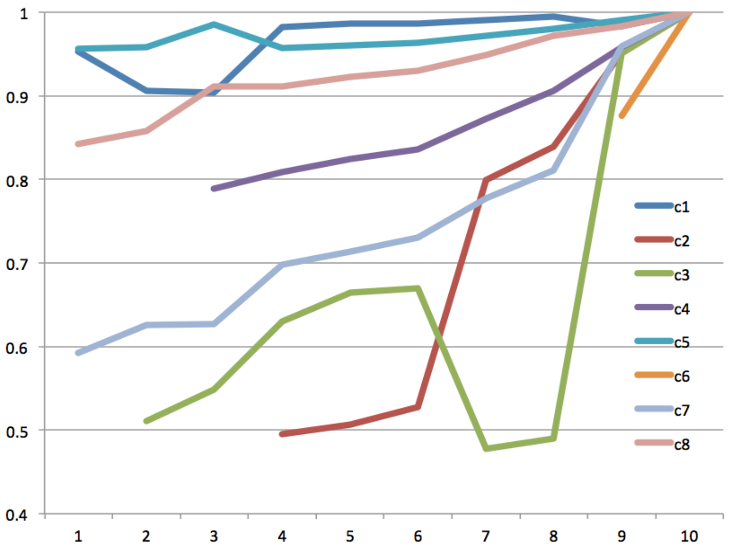

Figure 2: Cluster-dependent precision.

[2] The similarity threshold was set to Euclidian distance of *0.1*. This may be a basis for future experiments.

We note a considerable difference between the precision scores obtained by the clusters. A cross-cluster comparison, however, supports our previous finding: precision scores of clusters with a few dominant categories are higher than of those with uniform category scores. The highest precision was steadily achieved by clusters c1, c5, and c8 having one dominant category only. In fact, for the former two clusters the precision consistently hovered above the *0.9* mark starting from the first week. On the contrary, clusters c2 and c3 that had, respectively, three and four categories with scores higher than *0.1*, were the two worst performing clusters for the first seven weeks. Similarly, cluster c1 having the top category scoring *0.253* only, evolved after nine weeks and remained the worst performing clusters afterwards.

5 Conclusions

The plethora of accessible content in online catch-up TV services raises the emergent need for personalised recommendation solutions. In our work, we consider the use of stereotypical recommendation methods as the means to mitigate the cold-start problem, which is particularly acute in the catch-up TV scenario due to the continuous addition of new programs. The first step in this direction was to understand the ways to cluster users and obtain the stereotypical watching patterns characterising the identified clusters.

In this work we presented an exploratory study that applied clustering to group the users into eights clusters according to their observed watching behavior. We quantitatively and qualitatively analysed the identified clusters and pointed out a small number of program categories prevailing within each cluster. We also analysed the time-based stability of the clusters and the correlation between the stabilization time and the dominance of certain categories. The results allow us to conclude that clusters having a small number of dominant categories stabilise sooner and can be identified more accurately than others.

The next natural step of our work will be to conduct a small-scale user study aiming to validate that the identified clusters match stereotypical watching patterns of the population and to investigate whether the watching behavior in the clusters varies over time. Afterwards, we intend to use the identified clusters for the delivery of accurate and serendipitous recommendations to users. We will develop several offline recommendation methods and empirically evaluate their performance with the gathered offline dataset. The outcomes of this evaluation will inform the design of the catch-up TV recommendation service, which will be implemented and deployed in the future online evaluation with real users.

References

1. L. Ardissono, C. Gena, P. Torasso, F. Bellifemine, A. Diffno, and B. Negro. User Modeling and Recommendation Techniques for Personalized Electronic Program Guides. Personalized Digital Television, 2004.
2. P. Bellekens, G.J. Houben, L. Aroyo, K. Schaap, and A. Kaptein. User Model Elicitation and Enrichment for Context-Sensitive Personalization in a Multiplatform TV Environment. In Proceedings of the European Conference on Interactive Television, 2009.
3. R. Bernhaupt, M. Boutonnet, B. Gatellier, Y. Gimenez, C. Pouchepanadin and L. Souiba. A Set of Recommendations for the Control of IPTV-Systems via Smart Phones based on the Understanding of Users Practices and Needs. In Proceedings of the European Conference on Interactive Television, 2012
4. D. Bonnefoy, M. Bouzid, N. Lhuillier and K. Mercer. "More Like This" or "Not for Me": Delivering Personalised Recommendations in Multi-User Environments. In Proceedings of the International Conference on User Modeling, 2007.
5. J. Masthoff. Group modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers. User Modeling and User-Adapted Interaction, 14(1), 2004.
6. N. O'Connor, S. Marlow, N. Murphy, A. Smeaton, P. Browne, S. Deasy, H. Lee, and K. McDonald. Físchlár: an On-line System for Indexing and Browsing of Broadcast Television Content. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2001.
7. D. O'Sullivan, B. Smyth, D. Wilson, K. McDonald, and A. Smeaton. Interactive Television Personalization. In Personalized Digital Television, volume 6 of Human Computer Interaction Series, Springer 2004.
8. D. O'Sullivan, B. Smyth, D.C. Wilson, K. McDonald, and A.F. Smeaton. Improving the Quality of the Personalized Electronic Program Guide, User Modeling and User-Adapted Interaction, 14(1), 2004.
9. B. Smyth and P. Cotter. A Personalised TV Listings Service for the Digital TV Age. Journal of Knowledge-Based Systems, 13(2-3), 2000.
10. P.N. Tan, M. Steinback, and V. Kumar. Introduction to Data Mining: Pearson Addison Wesley, 2006.
11. TV and Video 2011 Consumer Trends Report, Ericsson ConsumerLab, 2011.
12. I. Witten, E. Frank and M. Hall. Data Mining: Practical Machine Learning Tools and Techniques: Morgan Kaufmann, 2011.
13. Z. Yu, X. Zhou, Y. Hao, and Jianhua Gu. TV Program Recommendation for Multiple Viewers Based on user Profile Merging. User Modeling and User-Adapted Interaction, 16 (1), 2006.
14. J. Zimmerman, K. Kauapati, A.L. Buczak, D. Schaffer, S. Gutta, and J. Martino. TV Personalization System. In Personalized Digital Television, volume 6 of Human Computer Interaction Series, Springer, 2004.