Time-Aware Evaluation of Methods for Identifying Active Household Members in Recommender Systems

Pedro G. Campos[1,2], Alejandro Bellogín[2], Iván Cantador[2], Fernando Díez[2]

[1] Departamento de Sistemas de Información
Universidad del Bío-Bío
4081112, Concepción, Chile
[2] Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
pgcampos@ubiobio.cl,
{pedro.campos, alejandro.bellogin, ivan.cantador,
fernando.diez}@uam.es

Abstract. Online services are usually accessed via household accounts. A household account is typically shared by various users who live in the same house. This represents a problem for providing personalized services, such as recommendation. Identifying the household members who are interacting with an online system (e.g. an on-demand video service) in a given moment, is thus an interesting challenge for the recommender systems research community. Previous work has shown that methods based on the analysis of temporal patterns of users are highly accurate in the above task when they use randomly sampled test data. However, such evaluation methodology may not properly deal with the evolution of the users' preferences and behavior through time. In this paper we evaluate several methods' performance using time-aware evaluation methodologies. Results from our experiments show that the discrimination power of different time features varies considerably, and moreover, the accuracy achieved by the methods can be heavily penalized when using a more realistic evaluation methodology.

Keywords: household member identification, time-aware evaluation, evaluation methodologies, recommender systems.

1 Introduction

Many online services providers offer access to their services via user accounts. These accounts can be seen as a mechanism to identify the active user, and track her behavior, letting e.g. build a personalized profile. A user profile can be used afterwards to provide personalized services, e.g. recommendation. However, user accounts can be shared by multiple users. An example of shared account is a household account, that is, an account shared by several users who usually live in

the same house. In general, it is hard to detect whether a user account is being accessed by more than one user, which raises difficulties for providing personalized services [1,2].

Users sharing a household do not necessarily access the service together. Consider for instance a four members family (formed e.g. by a father, a mother, a son and a daughter), sharing a household account of video-on-demand service. Each member of the family has distinct viewing interests and habits, and thus each of them watches video differently. If one member of the family asks for video recommendations, it is likely that those recommendations do not fit the user's interests, because the account profile contains a mixture of preferences from the four family members.

Two main strategies can be adopted in order to overcome such problem [3]. The first strategy is to increase the diversity of delivered recommendations [4], aiming to cover the heterogeneous range of preferences of the different members in a household. The second strategy is to identify the active household members for which recommendations have to be delivered. In this paper, we focus on the second strategy since it lets make more accurate recommendations, by only using preferences of active members, and discarding preferences of other, non-present members [1].

Previous work on the task has shown that the analysis of temporal patterns on historical data of household accounts provides important information for the discrimination of users, letting accurately identify active members [3,5,6]. Nonetheless, it is important to note that proposed methods have been assessed using evaluation methodologies based on the random selection of test cases. In a recent study on evaluation methodologies for recommender systems [7] it has been argued, however, that using randomly selected test data may not be fair for evaluation, particularly when temporal trends are being considered by the evaluated methods. We question whether this is also applicable for the task at hand, and in such case, which accuracy for active user identification would be achieved by using a more realistic evaluation methodology.

Using different evaluation methodologies, in this paper we perform an empirical comparison of methods for active household member identification in recommender systems. The tested methods are based on exploiting time information, and thus, we include some stricter time-aware evaluation methodologies. Results obtained from experiments on a real dataset show that the contribution of time features vary considerable when assessed by different methodologies, and moreover, the accuracy achieved by the methods can be heavily penalized when using a more realistic evaluation methodology.

The reminder of the paper is structured as follows. In Section 2 we describe related work. In Section 3 we detail methodologies employed in recommender systems evaluation that can be applied for assessing accuracy of methods used for identifying active household members. In Section 4 we present the methods evaluated. In Section 5 we describe the experiments performed, and report the results obtained. Finally, in Section 6 we present some conclusions and lines of future work.

2 Related work

The convenience of identifying users in households for recommendation purposes has been addressed in the recommender systems (RS) literature. Several proposals of RS on the TV domain consider the knowledge of which users are receiving the recommendations by means of explicit identification of users. For instance, Ardissono et al. [8] propose a personalized Electronic Programming Guide for TV shows, requiring the user to log in the system for providing personalization. Vildjiounaite et al. [9] propose a method to learn a joint model of users subsets in households, and use individual remote control devices for identifying users. The methods considered in this work, in contrast, aim to identify the user who is currently interacting with the system, by analyzing temporal patterns of individual users, without requiring to log in or to use special devices at recommendation time.

Specific methods for the identification of users from household accounts have been proposed in the RS research field. Goren-Bar and Glinansky [10] predict which users are watching TV based on a temporal profile manually stated. In [10] users indicate the time lapses in which they would probably be in front of the TV. Oh et al. [11] derive time-based profiles from household TV watching logs, which model preferences for viewing of time lapses instead of individual users. In this way, the target profile corresponds to the time lapse at which recommendations are requested. These methods assume that users have a fixed temporal behavior through time.

Recently, the 2011 edition of the Context-Aware Movie Recommendation (CAMRa) Challenge [2] requested participants to identify which members of particular households were responsible for a number of events –interactions with the system in the form of ratings. The contest provided a training dataset with information about ratings in a movie RS, including the household members who provided the ratings, and the associated timestamps. The challenge's goal was to identify the users who had been responsible for certain events (ratings), and whose household and timestamp were given in a randomly sampled test dataset. This task is assumed to be equivalent to the task of identifying active users requesting recommendations at a particular time.

The winners of the 2011 CAMRa challenge [6] and some other participants (e.g. [5,12]) exploited several time features derived from the available event timestamps. Such features showed different temporal rating habits of users in a household, regarding the day of the week, the hour of the day, and the absolute date when users rate items. In subsequent work [3], additional time features were investigated, as well as classification methods that enable an easy exploitation of such features, achieving a very high accuracy in the task (~ 98%). In this paper we use some of the best performing methods and time features presented in [3], and assess them using stricter evaluation methodologies, in order to test the reliability of the methods.

As a matter of fact, researchers in the RS field have questioned the suitability of some evaluation methodologies used for assessing RS that exploit temporal patterns in data [13,14]. Their main objection is that data used for test purposes is not always more recent than data used for training, and this may be unfair for methods exploiting time knowledge. In [7] we compared several RS using different evaluation

methodologies, and found that measured performance and relative ranking of methods may vary considerably among methodologies. Extrapolating findings obtained in that work to the task at hand, we can expect to find differences in the accuracy of methods by utilizing methodologies that do not use randomly sampled test data.

3 Evaluation Methodologies for Recommender Systems

The evaluation of recommender systems can be performed either *online* or *offline* [15]. In an *online evaluation* real users interactively test one or more deployed systems, and in general, empirical comparisons of user satisfaction for different item recommendations are conducted by means of A/B tests [16]. In an *offline evaluation*, on the other hand, past user behavior recorded in a database is used to assess the systems' performance, by testing whether recommendations match the users' declared interests. Given the need of having deployed systems and a large number of people using them in online evaluations, and the availability of historical users' data, most work in the RS field –and the one presented here– have focused on offline evaluations.

From a methodological point of view, offline evaluation admits diverse strategies for assessing RS performance. In general, a recommendation model is built (trained) with available user data, and afterwards its ability to deliver *good*[1] recommendations is assessed somehow with additional (test) user data. From this, in an offline evaluation scenario, we have to simulate the users' actions *after* receiving recommendations. This is achieved by splitting the set of available ratings into a *training set* –which serves as historical data to learn the users' preferences– and a *test set* –which is considered as knowledge about the users' decisions when faced with recommendations, and which is commonly referred to as *ground truth* data. As noted in [15], there are several ways to split data into training and test sets, and this is a source for differences in evaluation of RS. Moreover, in case that the data is time stamped –the case for RS exploiting time information– differences in evaluation can be meaningful, and may affect relative ranking of the algorithms' performance [7].

Several offline evaluation methodologies had been employed in measuring recommender system performance. In [7] several time-aware (and time-unaware) methodologies are described, by means of a methodological description framework that is based on a number of key methodological conditions that drive a RS evaluation process. These methodological conditions include: a) the *rating order* criterion ($\mathcal{O}$) used for split data. For instance, we may use a time-dependent ordering of data ($\mathcal{O}_{td}$), assigning the last (according to timestamp) data to the test set. Or, we may assign a random subset of data (i.e. time-independent ordering, $\mathcal{O}_{ti}$) to the test set; b) the *base set* ($\mathcal{B}$) on which the rating order criterion is applied. For instance, the ordering criterion can be applied on the whole dataset (a community centered base set, $\mathcal{B}_{cc}$), or can be applied independently over each user's data (a user-centered base set, $\mathcal{B}_{uc}$). In

[1] There is no general definition of what *good* recommendations are. Nonetheless, a commonly used approach is to establish the quality (goodness) of recommendations by computing different metrics that assess various desired characteristics of RS outputs.

the latter case, the last data from *each user* is assigned to the test set. This case, despite its popularity due to its application in the Netflix Prize competition [17], is not the best choice to mimic real-world evaluation conditions [7]; and c) the *size* condition ($\mathcal{s}$), i.e., the number of ratings selected for the test set. For instance, a proportion-based schema can be used ($\mathcal{s}_{prop}$), e.g. assigning 20% of data to the test set and the remaining 80% to the training set, or a fixed number q of ratings per user can be assigned to the test set ($\mathcal{s}_{fix,q}$), assigning the remaining ratings from each user to the training set.

The above evaluation conditions and related methodologies can be easily extrapolated to the task at hand, in order to test the reliability of the existent methods for household member identification. The only condition that requires a special treatment is the base set condition. In this case, as available data includes the household at which each user belongs to, it is possible to define a *household-centered* base set ($\mathcal{b}_{hc}$). That is, the application of rating order and size conditions on each household's data.

4 Methods for Identifying Active Household Members

Following the formulation given in [3], we treat the identification of the active household members as a classification problem, aiming to classify user patterns described by feature vectors that include time context information. This approach can be formalized as follows. Let us consider a set of events $E = \{e_1, e_2, \dots, e_m\}$ and a set of users $U_h = \{u_{1,h}, u_{2,h}, \dots, u_{n,h}\}$ within a household h, such that event e_i is associated to one, and only one, user $u_{j,h}$. Also, let us consider that each of these events is described by means of a feature vector, called X_{e_i}. The question to address is whether it is possible to determine which user is associated to an event e_i once (some) components x_{e_i} of its feature vector X_{e_i} are already known. In this paper, events correspond to instances of user ratings, and feature vectors correspond to time context representations of the events. Based on findings in [3], the time features considered in this work are the **absolute date** (D), the **day of the week** (W) and the **hour of the day** (H), as they are the best performing features reported in that paper for this task.

The first method considered is the *A priori* model described in [3]. This method computes probability distribution functions, which represent the probabilities that users are associated to particular events, and uses computed probabilities to assign a score to each user in a household given a new event. More specifically, we compute the probability mass function (PMF) of each feature given a particular user, restricted to the information related with that user's household, that is, $\{p(X = x_i|u_j)\}_{u_j \in U_h}$, where U_h is the set of users in the household h. Then, for each new event e, we obtain its representation as a feature vector $\hat{X}_e$, and identify the user who maximizes the PMF, that is, $u_j^*(e) = \arg\max_{u_j \in U_h} p(\hat{X}_e|u_j)$. When more than one feature is used, we assume independence and use the joint probability function, i.e., the product of the features' PMFs.

We also evaluate Machine Learning (ML) algorithms described in [3], that are able to deal with heterogeneous attributes. Specifically, we have restricted our study to the

following methods: Bayesian Networks (BN), Decision Trees (DT), and Logistic Regression (LR). These methods provide a score $\{s(\hat{X}_e, u_j)\}_{u_j \in U_h}$ based on different statistics from the training data, and select the users with highest scores.

The above methods use a fixed set of time features in the classification task, i.e., they use the same set of features over all the households. It is important to note, however, that data from only one household is used for classifying events of that household, i.e., the methods do not use data from other households for identifying members of a given household.

5 Experiments

In this section we report and discuss results obtained in experiments we conducted to evaluate the methods presented in Section 4, by means of different evaluation methodologies. Using some time-aware methodologies, we aim to test the reliability of the methods for identification of active household members in a realistic scenario. We begin by describing the used dataset, followed evaluation methodologies, and assessed metric.

5.1 Dataset

We use a real movie rating dataset made publicly available by MoviePilot[2] for the 2011 edition of the CAMRa Challenge [2]. This dataset contains a training set of 4,546,891 time stamped ratings from 171,670 users on 23,974 movies, in the timespan from July 11, 2009 to July 12, 2010. A subset of 145,069 ratings contains a household identifier. This subset includes a total of 602 users from 290 different households, who rated 7710 movies. The dataset also includes two test sets that also contain ratings with household identifier. Test set #1 contains 4482 ratings from 594 users on 811 items in the timespan from July 15, 2009 to July 10, 2010, and Test set #2 contains 5450 ratings from 592 users on 1706 items in the timespan from July 13, 2009 to July 11, 2010. We merged all the ratings with household identification, obtaining a total of 155,001 unique ratings (the *household dataset*). These ratings were then used for building several training and test sets according to different evaluation methodologies, as described below.

5.2 Evaluation Methodologies and Metrics

Aiming to analyze differences on accuracy of the methods presented in Section 3, we selected three different evaluation methodologies. Two of them use a time-dependent rating order condition, and the other one use a time-independent order condition.

The first methodology (denoted as $\mathscr{b}_{cc}\sigma_{td}\mathscr{s}_{fix}$) consists of a combination of a community-centered base set ($\mathscr{b}_{cc}$), a time-dependent rating order (σ_{td}), and a fixed

[2] www.moviepilot.com

size ($\mathcal{s}_{fix,q=5450}$) condition. Specifically, all ratings in the household dataset are sorted according to their timestamp, and the last 5,450 ratings are assigned to the test set (and the first 149,551 are assigned to the training set). In this way, a test set of similar size to test set #2 is built. The second methodology (denoted as $\mathcal{b}_{hc}\mathcal{o}_{td}\mathcal{s}_{fix}$) is equivalent to $\mathcal{b}_{cc}\mathcal{o}_{td}\mathcal{s}_{fix}$ with a household-centered base set condition ($\mathcal{b}_{hc}$). Specifically, the ratings of each household are sorted according to timestamp, and the last 19 ratings from each household are assigned to the test set. We chose 19 ratings aiming to build a test set of similar size to the one built with $\mathcal{b}_{cc}\mathcal{o}_{td}\mathcal{s}_{fix}$. The third methodology (denoted as $\mathcal{b}_{hc}\mathcal{o}_{ti}\mathcal{s}_{fix}$) is similar to $\mathcal{b}_{hc}\mathcal{o}_{td}\mathcal{s}_{fix}$ with a time-independent rating order condition ($\mathcal{o}_{ti}$). That is, 19 ratings are randomly selected from each household, and assigned to the test set.

We computed the accuracy of the evaluated methods in terms of the correct classification rate by household ($acc_{\mathbb{H}}$), i.e., the number of correct active member predictions divided by the total number of predictions, averaged by household, as proposed by CAMRa organizers. Formally, let $\mathbb{H}$ be the entire set of households in the dataset, and let $f(\cdot)$ be a method under evaluation. The metric is expressed as follows:

$$acc_{\mathbb{H}} = \frac{1}{\mathbb{H}} \sum_{h \in \mathbb{H}} \frac{1}{h} \sum_{(e_i,u_i) \in h} L\big(u_i, f(e_i)\big)$$

where $f(e_i) = \hat{u}$ is the user predicted by $f(\cdot)$ as associated to e_i, $L(u,\hat{u}) = 1$ if $u = \hat{u}$, and 0 otherwise, and (e_i, u_i) are the pairs of events and users of household h in the test set.

5.3 Results

Table 1 shows the $acc_{\mathbb{H}}$ results obtained by the evaluated methods using the three methodologies detailed in Section 5.2. The table also shows the results obtained on the test set #2, proposed by CAMRa organizers for the task (column titled CAMRa).. The table shows the results obtained by using individual time features, grouped by method.

In the table, we observe similar results when using methodologies based on a time-independent (random) rating order condition (CAMRa and $\mathcal{b}_{hc}\mathcal{o}_{ti}\mathcal{s}_{fix}$). Much worse results are observed when using methodologies employing a time-dependent rating order condition ($\mathcal{b}_{cc}\mathcal{o}_{td}\mathcal{s}_{fix}$ and $\mathcal{b}_{hc}\mathcal{o}_{td}\mathcal{s}_{fix}$). Particularly lower accuracies are achieved when using $\mathcal{b}_{cc}\mathcal{o}_{td}\mathcal{s}_{fix}$. We note that this methodology provides the evaluation scenario most similar to a real-world situation: data up to a certain point in time is available for training purposes, and data after that (unknown at that time) is then used as ground truth. In our case, this methodology provides a small number of training events for some households, which affect the methods' ability to detect temporal patterns of users. In fact, for some households, there is no training data at all. In this way, $\mathcal{b}_{cc}\mathcal{o}_{td}\mathcal{s}_{fix}$ represents a hard, but realistic evaluation methodology for the task. On the contrary, methodologies using a time-independent rating order condition provide easy, but unrealistic evaluation scenarios, because they let the methods use training data that would not be available in a real-world setting. The $\mathcal{b}_{hc}\mathcal{o}_{td}\mathcal{s}_{fix}$ method-

ology provides an intermediate scenario, in which an important part of data is available for learning temporal patterns of each household's members.

We also observe in Table 1 that the discrimination power of the different time features varies among methodologies. In the case of the A priori method, the best results on time-independent methodologies and $\mathscr{b}_{hc}\mathscr{o}_{td}\mathscr{s}_{fix}$ are obtained with **hour of the day** (H) feature, while **absolute date** (D) achieves the best results among ML methods –we note that results are similar across features. However, when using the stricter $\mathscr{b}_{cc}\mathscr{o}_{td}\mathscr{s}_{fix}$, the best results among methods are obtained with **day of the week** (W) feature, nearly followed by the **hour of the day** feature. On the contrary, the **absolute date** feature performs the worst consistently. This highlights how unrealistic the less strict methodologies are for the task, because they let the methods exploit a temporal behavior (the exact date of interaction) that in a real situation would be impossible to learn. This also shows that **hour of the day**, and more strongly **day of the week**, features describe a consistent temporal pattern of users through time.

Table 2 shows the $acc_{\mathbb{H}}$ results obtained by the evaluated methods using combinations of time features, and the same methodologies reported in Table 1. The results show that using less strict methodologies, combinations including the **absolute date** feature perform better. On the contrary, using $\mathscr{b}_{cc}\mathscr{o}_{td}\mathscr{s}_{fix}$ the best results are achieved by the combination of **hour of the day** and **day of week**.

All these results show that correct classification rate is prone to major differences depending on the evaluation methodology followed. The discrimination power of time features varies considerably when assessed by different methodologies. Moreover, the accuracy achieved by the methods is much lower when using the more realistic $\mathscr{b}_{cc}\mathscr{o}_{td}\mathscr{s}_{fix}$ methodology.

6 Conclusions and Future Work

In this paper we have presented an empirical comparison of methods for active household member identification, evaluated under different methodologies previously applied on recommender systems evaluation. Given that the methods are based on exploiting temporal patterns, we included some time-aware evaluation methodologies in order to test the reliability of previously reported results. We also analyzed the contribution of each time feature and combinations of features to the task.

The results obtained show that the discrimination power of time features, alone and combined, varies considerably when assessed by different methodologies. We observed that less strict methodologies provide unrealistic results, due to the exploitation of temporal information that are hard to obtain in a realistic evaluation scenario. Moreover, the accuracy achieved by all the methods was much worse when using a strict time-aware evaluation methodology. This findings show that stronger methods are required to provide accurate identification of active household members in real-world applications.

Next steps in our research will consider the development of methods able to improve accuracy in the task on the stricter time-aware evaluation methodologies, as a previous step towards obtaining better results on real-world applications. One way to

accomplish this goal may be to exploit patterns found across several households that may be useful to use in cases where little information about user's temporal behavior is available. Furthermore, we plan to test additional time features that can be derived from timestamps, and use a combination of time features and other type of features, e.g. based on demographic data, aiming to increase the discrimination power of the feature set.

Table 1. Correct classification rates obtained by the evaluated methods using the different time features and evaluation methodologies. Global top values in each column are in bold, and the best values for each method are underlined.

Method	Time Feature	$\mathscr{b}_{cc}\sigma_{td}\mathscr{s}_{fix}$	$\mathscr{b}_{hc}\sigma_{td}\mathscr{s}_{fix}$	$\mathscr{b}_{hc}\sigma_{ti}\mathscr{s}_{fix}$	CAMRa
A priori	*H*	0.6087	0.8163	0.9468	0.9457
	W	0.6167	0.8069	0.9299	0.9310
	D	0.4947	0.8152	0.9461	0.9413
BN	*H*	0.6533	0.8232	0.9539	0.9442
	W	0.6907	0.8189	0.9412	0.9438
	D	0.6506	**0.8575**	**0.9574**	**0.9538**
DT	*H*	0.6637	0.8229	0.9541	0.9459
	W	**0.6963**	0.8223	0.9417	0.9435
	D	0.6506	0.8544	0.9535	0.9472
LR	*H*	0.6674	0.8256	0.9537	0.9432
	W	0.6908	0.8132	0.9381	0.9405
	D	0.6147	0.8307	0.9555	0.9515

Table 2. Correct classification rates obtained by the evaluated methods using combinations of time features, on different evaluation methodologies. Global top values in each column are in bold, and best values for each method are underlined.

Method	Time Feature	$\mathscr{b}_{cc}\sigma_{td}\mathscr{s}_{fix}$	$\mathscr{b}_{hc}\sigma_{td}\mathscr{s}_{fix}$	$\mathscr{b}_{hc}\sigma_{ti}\mathscr{s}_{fix}$	CAMRa
A priori	*HW*	0.6496	0.8421	0.9688	0.9652
	HD	0.4947	0.8205	0.9739	0.9727
	WD	0.4947	0.8152	0.9470	0.9426
	HWD	0.4947	0.8205	0.9746	0.9720
BN	*HW*	0.6876	0.8325	0.9721	0.9690
	HD	0.6262	0.8287	0.9773	0.9740
	WD	0.6529	0.8127	0.9534	0.9484
	HWD	0.6809	0.8401	0.9770	0.9744
DT	*HW*	**0.7188**	0.8644	0.9773	0.9750
	HD	0.6389	0.8648	0.9753	0.9709
	WD	0.6932	0.8417	0.9526	0.9470
	HWD	0.6950	0.8599	0.9777	0.9752
LR	*HW*	0.6635	0.8652	0.9768	0.9701
	HD	0.6515	0.8650	**0.9824**	**0.9769**
	WD	0.6636	**0.8697**	0.9553	0.9564
	HWD	0.6591	0.8670	0.9808	0.9759

References

1. Kabutoya, Y., Iwata, T., Fujimura, K.: Modeling Multiple Users' Purchase over a Single Account for Collaborative Filtering. In: *Proceedings of the 11th International Conference on Web Information Systems Engineering*, pp. 328–341 (2010)
2. Berkovsky, S., Luca, E.W. De, Said, A.: Challenge on Context-Aware Movie Recommendation: CAMRa2011. In: *Proceedings of the 5th ACM Conference on Recommender Systems*, pp. 385–386 (2011)
3. Campos, P.G., Bellogin, A., Díez, F., Cantador, I.: Time feature Selection for Identifying Active Household Members. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2311–2314 (2012)
4. Zhang, M., Hurley, N.: Avoiding Monotony: Improving the Diversity of Recommendation Lists. In: *Proceedings of the 2nd ACM Conference on Recommender Systems*, pp. 123–130 (2008)
5. Campos, P.G., Díez, F., Bellogín, A.: Temporal Rating Habits: A Valuable Tool for Rater Differentiation. In: *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, pp. 29–35 (2011)
6. Bento, J., Fawaz, N., Montanari, A., Ioannidis, S.: Identifying Users from Their Rating Patterns. In: *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, pp. 39–46 (2011)
7. Campos, P.G., Díez, F., Cantador, I.: Time-Aware Recommender Systems: A Comprehensive Survey and Analysis of Existing Evaluation Protocols. *User Modeling and User-Adapted Interaction*. In press (2013)
8. Ardissono, L., Portis, F., Torasso, P., Bellifemine, F., Chiarotto, A., Difino, A.: Architecture of a System for the Generation of Personalized Electronic Program Guides. In: *Proceedings of the UM'01 Workshop on Personalization in Future TV* (2001)
9. Vildjiounaite, E., Hannula, T., Alahuhta, P.: Unobtrusive Dynamic Modelling of TV Program Preferences in a Household. In: *Proceedings of the 6th European Conference on Changing Television Environments*, pp. 82–91 (2008)
10. Goren-Bar, D., Glinansky, O.: FIT-recommending TV Programs to Family Members. *Computers & Graphics* 24, pp. 149–156 (2004)
11. Oh, J., Sung, Y., Kim, J., Humayoun, M., Park, Y.-H., Yu, H.: Time-Dependent User Profiling for TV Recommendation. In: *Proceedings of the 2nd International Conference on Cloud and Green Computing*, pp. 783–787 (2012)
12. Shi, Y., Larson, M., Hanjalic, A.: Mining Relational Context-aware Graph for Rater Identification. In: *Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation*, pp. 53–59 (2011)
13. Lathia, N., Hailes, S., Capra, L.: Temporal Collaborative Filtering with Adaptive Neighbourhoods. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 796–797 (2009)
14. Campos, P.G., Díez, F., Sánchez-Montañés, M.: Towards a More Realistic Evaluation: Testing the Ability to Predict Future Tastes of Matrix Factorization-based Recommenders. In: *Proceedings of the 5th ACM conference on Recommender systems*, pp. 309–312 (2011)
15. Shani, G., Gunawardana, A.: Evaluating Recommendation Systems. In: Ricci, F., Rokach, L., Shapira, B., and Kantor, P.B. (eds.) *Recommender Systems Handbook*, pp. 257–297. Springer US (2011)
16. Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M.: Controlled Experiments on the Web: Survey and Practical Guide. *Data Mining and Knowledge Discovery* 18, pp. 140–181 (2008)
17. Bennett, J., Lanning, S.: The Netflix Prize. In: *Proceedings of KDD Cup and Workshop 2007* (2007)