A Dataset for Inferring Contextual Preferences of Users Watching TV

Kristoffersen, Miklas Strøm; Shepstone, Sven Ewan; Tan, Zheng-Hua

Published in:
UMAP 2018 - Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization

DOI (link to publication from Publisher):
10.1145/3209219.3209263

Publication date:
2018

A Dataset for Inferring Contextual Preferences of Users Watching TV

Miklas S. Kristoffersen
Bang & Olufsen A/S & Aalborg Univ.
Struer, Denmark
mko@bang-olufsen.dk

Sven E. Shepstone
Bang & Olufsen A/S
Struer, Denmark
ssh@bang-olufsen.dk

Zheng-Hua Tan
Aalborg University
Aalborg, Denmark
zt@es.aau.dk

ABSTRACT

Studies have shown that contextual settings play an important role in users' decision processes of what to consume, but data supporting the investigation of context-aware recommender systems are scarce. In this paper we present a TV consumption dataset enriched with contextual information of viewing situations. The dataset is designed for studying the intrinsic complexity of TV watching activities, and hence we also evaluate the performance of predicting chosen genres given contextual settings, and compare the results to contextless predictions. The results suggest a significant improvement by including contextual features in the prediction.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing → User studies**;

KEYWORDS

Context-Awareness; Recommender Systems; Machine Learning; Experience Sampling; Dataset; Television.

ACM Reference Format:
Miklas S. Kristoffersen, Sven E. Shepstone, and Zheng-Hua Tan. 2018. A Dataset for Inferring Contextual Preferences of Users Watching TV. In *UMAP '18: 26th Conference on User Modeling, Adaptation and Personalization, July 8–11, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3209219.3209263

1 INTRODUCTION

The concept of context-aware recommendations has been studied in several academic and commercial projects [2, 9], but there is still a need for publicly available datasets since only a limited number of such datasets exist, e.g. [1, 6, 10]. Also, temporal context has constituted a significant part of development and evaluation within context-aware recommender systems (CARS), since timestamps are often logged together with events, e.g. ratings, which allows for a simple way to reformulate challenges designed for traditional recommender systems into the CARS domain by using timestamps as temporal context. However, previous studies of users' TV watching behavior in given contexts have shown that the TV is mostly a social platform and consumption takes place in a wide variety of situations [8]. Furthermore, it has been shown that both temporal and social settings are key contextual indicators of what content is consumed [5, 11]. It is, however, challenging to collect TV consumption data that includes contextual information beyond timestamps, such as social settings. People meters, for instance, are challenged, [3], by non-compliance (participants neglect to push a button), and secondly, since meters log the opportunity to consume some content, there is no information of the actual exposure, i.e. the TV could be showing some content that the user does not watch.

In this work, we collect and analyze self-reported TV consumption data using the Experience-Sampling Method (ESM) [4]. We structure the data to accommodate quantitative analyses, e.g. in the CARS community. Lastly, using well-established Machine Learning methods we show performance in predicting consumed content given contextual settings, and compare this with contextless prediction.

2 CONTEXTUAL TV WATCHING DATASET

To obtain data from participants we developed a web page, and asked participants to answer questions five times every day at 8, 12, 17, 20, and 22 (or when going to bed) for a five week period. These intervals were chosen to accommodate work and study schedules, while still providing ample opportunity to participate over a full day period. Participants were allowed to answer more frequently (and at other times) than the five pre-specified intervals. To remind participants when to answer, we used a public calendar with alerts for iOS devices and web push notifications for all other types of devices.

Table 1: Questions and selection options in the dataset.

Questions		Options
Q1:	Have you watched TV within the last four hours?	Yes, no
Q2:	Who were you watching it with?	*Multiple-option:* Alone, partner, child (0-12), child (12+), sibling, parent, friend, other (text)
Q3:	How many people (including yourself) watched TV?	1, 2, 3, 4, 5+
Q4:	What did you watch?	*Multiple-option:* News, sport, movie, series, music, documentary, entertainment, children's, user-generated, other (text)
Q5:	Which service(s) did you use?	*Multiple-option:* Traditional TV, DRTV, TV2 Play, Viaplay, Netflix, HBO Nordic, YouTube, other (text)
Q6:	How much attention did you pay to the TV?	None-full (5 steps)

UMAP '18, July 8–11, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5589-6/18/07.
https://doi.org/10.1145/3209219.3209263

We collect the following background information the first time a participant login: gender, age group, language, device type, household size, additional household members, frequency of TV watching, and favorite TV genres. On subsequent logins, participants are asked the questions listed in Table 1. The general flow is that Q2-Q6 are asked only if the selection for Q1 is *yes*. Also, Q3 is skipped if *alone* is selected for Q2. For Q5 all except *Traditional TV* (and possibly *other*) are streaming services, some specific to Denmark/Scandinavia. The multiple-option questions allow more than one selection, e.g. *partner* and *friend*. Participants are instructed to split answers with different contextual settings, e.g. watching news alone and children's TV with a child. Answers are logged with the following format: Answer ID, User ID, timestamp, Q1, Q2, Q3, Q4, Q5, Q6. For further information see the publicly available dataset consisting of 118 participants and more than 6000 answers.[1]

3 EXPERIMENT

The goal of the experiment is to predict what genre a user is going to watch (Q4) in the reported context. The task is defined as a multi-class classification problem with the users' selections for Q4 as target. The selections for the remaining questions are used as contextual features. All features are categorical and represented using one-hot encoding. The optional text input for *other* in Q2, Q4, and Q5 are not included in this study.

The experiment includes two methods based on contextless (CL) and context-aware (CA) predictions, respectively. CL only takes the user ID of the respondent into account, while CA includes all the collected contextual features. A scikit-learn [7] implementation of logistic regression (LR) is used. We fit the LR weights using stochastic average gradient descent with L2 regularization, and set the multi-class parameter to "multinomial" for 10-way softmax regression. We also include two baseline methods for comparison, namely random and toppop. The random predictor randomly ranks the genres for each prediction. For toppop, genres are ranked by their popularity judged by the number of observations in the training set.

The methods are evaluated using nested cross-validation with five outer folds and three inner folds. That is, the training data for each outer fold are divided into three inner folds for optimization of hyperparameters. We report the average performance across the outer folds and the standard deviation. Performance is measured in terms of accuracy at K predictions (A@K)[2], F1 (macro), and mean reciprocal rank (MRR).

The results are shown in Table 2. Note that toppop outperforms random in terms of A@1, A@3, and MRR, but random performs better than toppop for F1 (macro), due to the diversity in predicted genres. The LR-based methods achieve considerably higher scores than both baseline methods, and furthermore CA-LR outperforms CL-LR. The MRR of CA-LR indicates that on average the true genre is ranked among the first and second (as indicated by 1/MRR≈1.5) of the 10 possible genres. The corresponding number for CL-LR is 1.7. McNemar's test[3] shows statistical significant improvement ($\chi^2(1)$=146.92, p<0.001, V=0.22) between the two methods.

Table 2: Results for the genre predictions (standard deviation in parentheses).

Method	A@1	A@3	F1 (macro)	MRR
random	0.093 (0.010)	0.296 (0.005)	0.083 (0.011)	0.289 (0.008)
toppop	0.245 (0.009)	0.560 (0.014)	0.039 (0.001)	0.460 (0.008)
CL-LR	0.368 (0.026)	0.761 (0.008)	0.244 (0.023)	0.586 (0.016)
CA-LR	0.487 (0.005)	0.849 (0.005)	0.446 (0.025)	0.679 (0.005)

4 CONCLUDING REMARKS

In this paper, an extensive field study over a period of five weeks with a group of more than 100 participants is used to evaluate to which degree contextual knowledge influences the performance of predicting what content will be consumed. The experimental results show that inclusion of contextual information significantly improves accuracy compared to contextless predictions.

In future work, we will apply state-of-the-art CARS methods to the dataset, and investigate the contribution of each contextual dimension.

ACKNOWLEDGMENTS

This work is supported by Bang and Olufsen A/S and the Innovation Fund Denmark (IFD) under File No. 5189-00009B.

REFERENCES

[1] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. 2005. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Trans. Inf. Syst.* 23, 1 (2005), 103–145.
[2] G. Adomavicius and A. Tuzhilin. 2015. Context-Aware Recommender Systems. In *Recommender Systems Handbook*. Springer, 191–226.
[3] B. Jardine, J. Romaniuk, J. G. Dawes, and V. Beal. 2016. Retaining the primetime television audience. *European Journal of Marketing* 50, 7/8 (2016), 1290–1307.
[4] R. Larson and M. Csikszentmihalyi. 1983. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science* (1983).
[5] K. Mercer, A. May, and V. Mitchel. 2014. Designing for video: Investigating the contextual cues within viewing situations. *Personal and Ubiquitous Computing* 18, 3 (01 Mar 2014), 723–735.
[6] C. Ono, Y. Takishima, Y. Motomura, and H. Asoh. 2009. Context-Aware Preference Model Based on a Study of Difference between Real and Supposed Situation Data. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization (UMAP '09)*. Springer Berlin Heidelberg, 102–113.
[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[8] D. Saxbe, A. Graesch, and M. Alvik. 2011. Television as a Social or Solo Activity: Understanding Families' Everyday Television Viewing Patterns. *Communication Research Reports* 28, 2 (2011), 180–189.
[9] Y. Shi, M. Larson, and A. Hanjalic. 2014. Collaborative Filtering Beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.* 47, 1, Article 3 (May 2014), 3:1–3:45 pages.
[10] R. Turrin, A. Condorelli, P. Cremonesi, and R. Pagano. 2014. Time-based TV programs prediction. In *1st Workshop on Recommender Systems for Television and Online Video at ACM RecSys (RecSys '14)*.
[11] J. Vanattenhoven and D. Geerts. 2015. Contextual aspects of typical viewing situations: a new perspective for recommending television and video content. *Personal and Ubiquitous Computing* 19, 5 (2015), 761–779.

[1] Available at http://kom.aau.dk/~zt/online/ContextualTVDataset.

[2] At K larger than one, multiple guesses are allowed for each test sample. It is calculated using: A@K = $\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{1}\left(\hat{y}_{n,k}=y_n\right)$, where N is the number of tests. $\mathbf{1}$ is the indicator function, which is one if the prediction, $\hat{y}_{n,k}$, is equal to the actual target, y_n, and zero otherwise.

[3] A matrix, $\boldsymbol{A}_{2\times2}$, is formed with $a_{1,1}$ being the number of tests where both methods are correct, $a_{1,2}$ and $a_{2,1}$ are the tests where one of the methods fail, and $a_{2,2}$ is when both are incorrect. McNemar's χ^2 test statistic and Cramér's V are then computed as: $\chi^2 = \left(a_{1,2}-a_{2,1}\right)^2/\left(a_{1,2}+a_{2,1}\right)$, $V=\sqrt{\chi^2/\sum_{i,j}a_{i,j}}$.