ELEKTROTEHNIŠKI VESTNIK 78(5): 270–274, 2011
ENGLISH EDITION

Database for contextual personalization

Andrej Košir[1], **Ante Odić**[1], **Matevž Kunaver**[1], **Marko Tkalčič**[1], **Jurij F. Tasič**[1]

[1]*Faculty of Electrical Engineering, University of Ljubljana,*
Tržaška 25, 1000 Ljubljana, Slovenia

† *E-mail: andrej.kosir@fe.uni-lj.si*

Abstract. In recent years, research into user centric and personalized applications has focused on the utilization of contextual information about the situation in which the user is consuming the content item. However, there is no database suitable for the investigation of specific open issues of contextual information description and utilization available today. The reason for this are several known difficulties with user related contextual information acquisition. A description of a publicly available database for personalization and user adaptation including contextual information is given in the paper. The data was acquired from users watching movies and then providing contextual information about the event in addition to submitting ratings about the movie. Beside describing raw data, the paper outlines basic statistics and selected properties of potentially contextual variables. At the time of submission, the database included 12 contextual variables, more than 90 users, 950 items and 1600 ratings. Data was acquired using a dedicated web application which is still publicly available and the data acquisition is still in progress. Content items (movies) can be enhanced by content item metadata using publicly available databases.

Keywords: Contextual personalization, test set, experimental design

1 INTRODUCTION

The problem of a huge and constantly growing number of available multimedia items and communication services is still unsolved from the user's point of view. The task of managing modern communication systems and their content is still not a comfortable or even feasible task when applied to a large segment of users. A major approach used to remedy these issues is personalization and user adaptation which is in large part based on the prediction of user decisions, especially ratings he/she would assign to content items.

Contextual personalization has received much attention in the last few years due to the fact that, among other factors, the context affects the way the user is consuming the given content item and it affects the decisions he/she makes about it. Context information is any information about the situation, circumstances and user state when a user is consuming the content item [1]. Context can be the time of day, weather, social situation, users mood etc. [2], [3]. However, the relationship between context and user decision making is very complex and difficult to model. In addition the contextual information is extremely difficult to acquire since the acquisition process usually interferes with the decision making process and may modify or even destroy the acquired context and consequently corrupt the acquired data.

Received December 9, 2011
Accepted December 16, 2011

A favorable approach to acquire user context data is to bring the data acquisition process as close to the user as possible. Movies are very popular content items and are frequently watched using personal computers, at least by younger users. Therefore, we have decided to use a web based application to acquire context data from users about the time and situation in which they consume the content item (watch the movie).

A contribution of our contextual dataset is context information acquired at the consumption stage and is therefore more reliable. In addition we captured 12 different types of (potentially) contextual information that allows the investigation of several different issues regarding contextual personalization. As indicated later, the *a priori* statistical power analysis showed that the required number of users, items and ratings is meet by our dataset.

This paper presents the dataset for contextual personalization development. Basic statistics and selected properties of included variables are given. Selected distinguished properties related to open issues of the presented dataset are discussed. The data acquisition process is still under way and we believe that our database provides the real consumption stage context [4] which is crucial for several open issues of context personalization. A discussion of possible future database development concludes the paper.

2 CONTEXTUAL PERSONALIZATION DATABASE

The acquisition of reliable and accessible contextual database has become crucial in the development of personalization algorithms. It is believed that the context has a significant impact on the user decision-making process and the presented data set allows the study of this impact. We present several issues relevant in the establishment of such a database. This section summarizes the general guidelines that are to be followed when collecting contextual personalization dataset. Details about our approach and the described dataset are given in the next section.

Regarding terminology, we use the term ”user data acquisition process” for the whole process of user data collection into the dataset. The term ”data acquisition application” is used to refer to the additional functions (radio buttons to provide feedback etc) added to the users communication device (PC used to watch movies, etc) in order to acquire his feedback.

2.1 How to select users and motivate them for contextual data acquisition?

As already indicated, the contextual data acquisition should not interfere with the user's decision making process. In the worst case this interference should be kept at the lowest level possible. This means that the user data acquisition process should be naturally embedded into the user's consumption and decision making environment.

It is also important that the user's collaboration should be driven by the right motives. We believe that the best motive is to help improve personalized applications as an end in itself and for other users [5].

To summarize, while providing contextual data and his decisions on content item consumption, the user should adhere to his/her usual ways and the circumstances in which he/she typically consumes this type of item. For instance, if the user usually watches movies using his personal computer (PC), the user data acquisition procedure should be offered to him via the same PC and with the least possible intrusion. Therefore, to build the dataset, we chose those users who already used communication devices that allowed the implementation and use of our user data acquisition application.

2.2 Which content items to include?

In the same way that the user context should be undisturbed during user data acquisition, the selection of content items offers to users during this acquisition should be the same as the selection otherwise available. The best solution for this issue is to make sure that the list of available items and the service providing these items is not changed when the user's data acquisition is in progress. As a consequence, the user data acquisition application should be independent of the content items of the service provider application.

2.3 How the data should be collected?

To achieve non-intrusive user data acquisition, data should be collected using communication devices and services that are used in real world applications. The additional functions required to collect the data should be easy to use, user-friendly (”gamified”) and, most importantly, as invisible to the user as possible. The perfect solution in which the user would not be aware of the data acquisition process is unfortunately not feasible today.

To come as close as possible to the perfect solution, the data acquisition application should be embedded into the users consumption device in a way that does not interfere with or destroys his/her usual patterns of behavior while consuming content items. These behavior patterns should not be changed by the data acquisition process when, for instance, the user is providing contextual data while in the company of his/her friends or family.

2.4 What additional information is the required to make the data set useful?

In order for the dataset to be applicable in contextual personalization research, raw user, item and context-related data should be equipped with the additional information. Basic information on users such as age, sex, content item related behavior patterns information (frequency of use, ...) should be provided in order to allow the statistical verification of dataset quality. Item metadata should also be provided or at least the way to provide it later should be secured such as movie metadata title, genre, actors, etc that can be retrieved from IMDB [6].

2.5 Which experimental designs should be supported?

Procedures for an exact evaluation of the efficiency of personalization algorithms are mostly based on statistical methods such as estimation of confusion matrix, estimation of ROC curve and statistical significance hypotheses testing. To use them properly, a suitable experimental design must be selected. Such an experimental design must be applicable to a dataset under investigation. This mainly depends on acquired variables and their properties related to single variables or a subset of variables. Firstly, the type of each variable, i.e. categorical, ordinal or numerical determines the types of applicable procedures. The size of the dataset is also decisive. A priori power analysis [7] for each experimental design can be used to verify whether the number of entries is sufficient for the planned interpretation of results. These numbers may refer to the number of users, items and ratings provided, to the density of the dataset (number of ratings per item etc).

Further details of experimental designs dependent on specific issues we wish to resolve. Clearly, not all requirements can be met. It is even more important that the dataset allows us to verify whether a given requirement is met or not.

3 CONSUMPTION CONTEXT MOVIES LDOS (LDOS-COMODA) CONTEXTUAL PERSONALIZATION DATASET

In this section, we describe the LDOS-CoMoDa dataset. The procedure of data acquisition is briefly outlined, available variables together with basic properties and statistics are given in order to allow the easy and efficient use of the dataset. Content items are movies, while the item consumption device is a personal computer with web-based application to acquire the user's context. Detailed information is given below.

3.1 Users, items, contextual variables and metadata

The LDOS-CoMoDa database has been created to meet requirements listed in the previous section as accuratelly as possible. It contains 30 variables among which are 12 contextual variables. Other variables are general user information (age, sex, city, country) and content (movie) metadata (director, movieCountry, movieLanguage, movieYear, genre1, genre2, genre3, actor1, actor2, actor3, budget).

Since contextual variables are of special interest to us, we describe them in Table 1. In terms of statistical classification, all of them are of the categorical or ordinal type. Some are categorical by nature (weather, etc.) and some were designed to be categorical or ordinal. The reason for this is twofold. The first is to control the number of classes we wish to cover in the process of the user data acquisition process and the second one is to simplify the analysis when selected issues of context personalization are investigated.

The LDOS-CoMoDa test set basic statistics at the moment of submission (15.12.2011) are given in Table 2.

When a generalization of results of statistics-based procedures is needed, the representativeness of the analyzed sample is of crucial importance. It is well known that the age of a user is an important factor in any prediction of his behavior related to modern communication devices. The histogram of user age is given in Figure 1. Distribution of ratings and users are given in Figure 2 and 3, respectively. Observe the long tails of these two ordered histograms.

The density of the database is depicted in Figure 4. The brighteness is proportional to the number of ratings assigned to a given subgroup of items (one column for each group of items) by a single user (one row for each user). A relatively high variability among users can be indicated. The same is true for items. Since user ids are added sequentially, one can observe that users who have joined later have contributed fewer ratings.

Var. Name	|Rg|	MVR	Description
time	4	0.017	morning, afternoon, evening, night
daytype	3	0.015	working day, weekend, holiday
season	4	0.017	spring, summer, autumn, winter
location	3	0.016	home, public place, friend's house
weather	5	0.021	sunny/clear, rainy, stormy, snowy, cloudy
social	7	0.013	alone, partner, friends, colleagues, parents, public, family
endEmo	7	0	sad, happy, scared, surprised, angry, disgusted, neutral
dominantEmo	7	0	sad, happy, scared, surprised, angry, disgusted, neutral
mood	3	0	positive, neutral, negative
physical	2	0.022	healthy, ill
decision	2	0.021	user's choice, given by other
interaction	2	0.020	first, n-th

Table 1. Contextual variables and their basic properties. |Rg| stands for the number of categorical or ordinal classes for the variable (size of its rank) and MVR stands for missing value ratio

Table 2. Basic dataset statistic

Number of users	95
Number of items	961
Number of ratings	1665
Average age	27.0
Number of countries	6
Number of cities	18
Max ratings of single user	220
Min ratings of single user	1

3.2 Data acquisition process

The data acquisition procedure for user contextual data is very sensitive to any context disturbance. It makes the correct interpretation of the test results obtained by using the dataset difficult. As indicated in Section 2, the data acquisition process may influence the acquired user context and interfere with his/her decision making process.

The data contained in the LDOS-CoMoDa dataset was acquired via a specially designed user friendly web application. It is important to note that users were instructed to provide ratings and context immediately after the consumption stage (watching the selected movie). It is believed that such a context is more reliable compared to the one captured later. In this way, the consump-

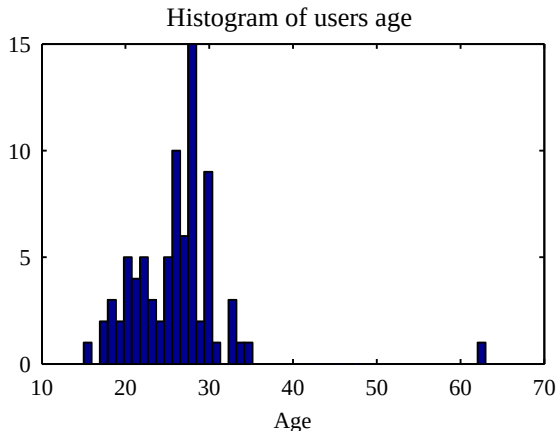

Figure 1. Histogram of user's age. Most users are aged between 18 and 35 what limits the representativeness of the dataset to this age interval.

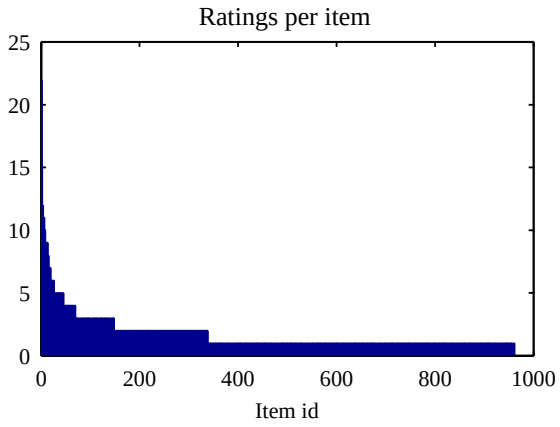

Figure 2. Number of ratings per item. Most of items received 2 or 3 ratings.

tion stage context is recorded including user emotions. Users were motivated to provide their feedback to help themselves and others improve their user model [5] and consequently the quality of personalized service. In addition, the acquisition application offers a personal tracking of movies watched, a search tool for selected context such as the time of day and collaborative filter recommendations as an additional benefit to those willing to contribute their rating and context.

Figure 5 shows the course of the number of users, items and ratings provided over time. It may be seen that the rate of growth is decreasing which is expected, and some local faster growths is clearly identifiable. The application is still active and accessible at http://212.235.187.145/spletnastran/raziskave/um/emotions/login.php.

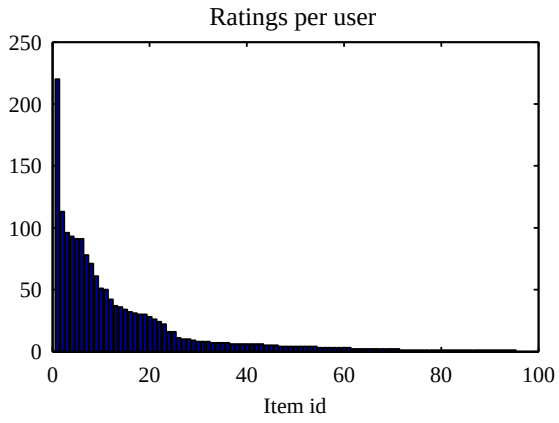

Figure 3. Number of ratings per user. We observe a high variability in the activity of users, from a large number of users providing fewer than 10 items to a large group of user with more than 50 ratings.

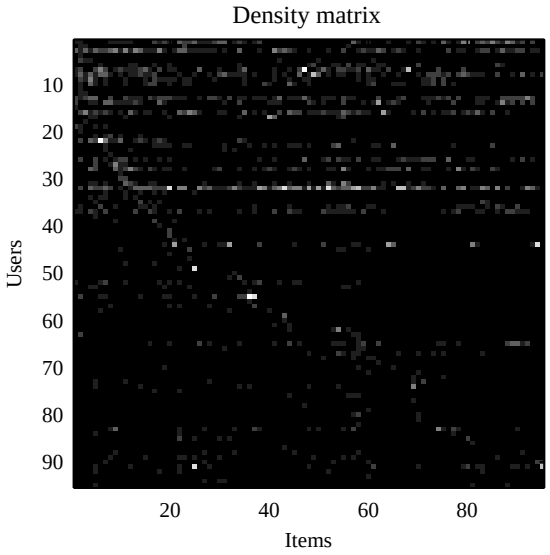

Figure 4. The density of the database. The brightnes is proportional to number of given ratings.

3.3 LDOS-CoMoDa dataset accessibility

In order to allow easy web access to the dataset and to follow the use of it by the research community, a password protected web link to the dataset will be made available after an e-mail has been sent to ldos-comoda@ldos.si. In addition to the dataset, the access instructions will be made available together with the updated version of this paper.

4 CONCLUSION

With more than 90 users, 900 items and 1600 ratings the LDOS-CoMoDa dataset provides a dataset of con-

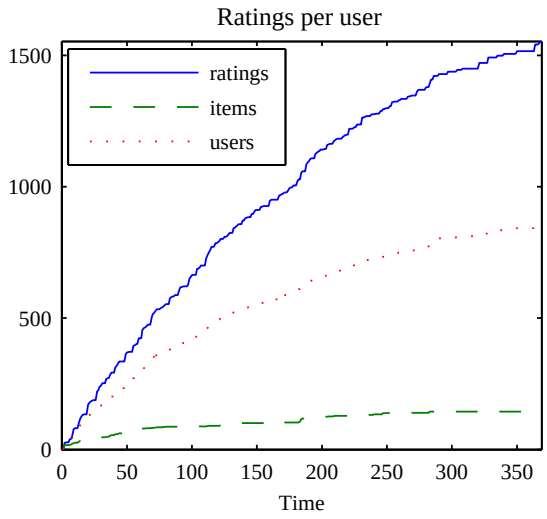

Figure 5. A course of test set data acquisition over time

sumption time and, contextual information, suitable for the investigation of several open issues in contextual personalization [4], [8], [9]. We believe that the most sensitive part of the database are the contextual variables describing the user emotions etc., which are relatively accurate descriptions of the users real context at the time of consumption (consumption stage). The database is publicly available and easy to access on prior e-mail request.

The most relevant advantage of our dataset is that it contains consumption stage context information which is believed to be more reliable. It includes 12 types of potentially contextual information. According to a priori statistical power analysis, the size and density of our dataset is sufficient to investigate several contextual personalization issues.

Nevertheless, the dataset has several limitations at this stage which is also the reason why the acquisition process is still active. *a priori* power analysis [7] (not reported in this paper) showed that at the typical effect size the required number of ratings is around 1400. Our dataset passed this limit but there are several interesting experiments where these ratings must be divided into subgroups where this limit is not met anymore. Larger numbers of users and their ratings should also be secured in order to improve the density of the dataset.

REFERENCES

[1] A. Dey, G. Abowd, Towards a better understanding of context and context-awareness, Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing (1999) 304–307.

[2] L. Baltrunas, B. Ludwig, S. Peer, F. Ricci, Context relevance assessment and exploitation in mobile recommender systems, Personal and Ubiquitous Computing (2011) 1–20doi:10.1007/s00779-011-0417-x.

[3] F. Díez, J. E. Chavarriaga, P. G. Campos, A. Bellogín, Movie Recommendations based in explicit and implicit features extracted from the Filmtipset dataset, in: Proceedings of the Workshop on Context-Aware Movie Recommendation, 2010, pp. 45—-52.

[4] A. Odić, M. Kunaver, J. Tasič, A. Košir, Open issues with contextual information in existing recommender system databases, in: ERK 2010 Proceedings, 2010.

[5] J. Herlocker, J. Konstan, L. Terveen, J. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems 22 (1) (2004) 5–53. doi:/10.1145/963770.963772.

[6] The internet movie database (imdb) @ONLINE (Dec. 2011). URL http://www.imdb.com/

[7] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum, 1988.

[8] G. Adomavicius, R. Sankaranarayanan, S. Sen, A. Tuzhilin, Incorporating contextual information in recommender systems using a multidimensional approach, ACM Transactions on Information Systems (TOIS) 23 (1) (2005) 103–145.

[9] Z. Yujie, W. Licai, Some Challenges for Context-aware Recommender Systems, in: Computer Science and Education (ICCSE), 2010 5th International Conference on, 2010, pp. 362–365.

Andrej Košir is an associate professor in the Faculty of Electrical Engineering at the University of Ljubljana. His research interests include operational research in telecommunication, user modeling, user adaptation and social signal processing.

Ante Odić is a junior researcher in the Faculty of Electrical Engineering at the University of Ljubljana. For his PhD, he is doing research on the use of contextual information in personalized services.

Matevž Kunaver is a part-time researcher and part-time teaching assistant in the Faculty of Electrical Engineering at the University of Ljubljana. He specializes in collaborative recommender systems as well as hybrid recommender systems incorporating techniques from content-based recommender systems.

Marko Tkalčič is a researcher in the Faculty of Electrical Engineering at the University of Ljubljana. He is currently doing research on the use of affective and personality parameters for modeling users and content in various applications.

Jurij F. Tasič is a full professor of system theory and computing at the University of Ljubljana. His current interests are advanced algorithms in communication systems, multidimensional signal processing and parallel algorithms.