User Modelling Focused on Short-Term Behavioural Changes

Ondrej Kassak*
Institute of Informatics, Information Systems and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovicova 2, 842 16 Bratislava, Slovakia
name.surname@stuba.sk

Abstract
Understanding and modelling of the short-term user behaviour represents one of the most relevant topics researched within the Web in last years. It provides information allowing to react to user actions quickly and in this way to increase the quality of user's interaction. Beside the traditional applications of user models as the personalization or recommendation, which depend mostly on stable long-term preferences, novel tasks as the session end intent or next action prediction are researched these days. These tasks depend on highly actual information allowing to react in an online time. Additionally, a type of the modelled data evolve, as the traditional preferences are often insufficient nowadays. In our thesis, we propose a novel user model capturing changes in short-term data on the level of session actions. We evaluate the model by the session end intent prediction task. To prove model characteristics, the evaluation was realised on the data from two domains with highly different characteristics.

Categories and Subject Descriptors
[**Human-centered computing**]: Human computer interaction (HCI) - HCI design and evaluation methods - User models; [**Information systems**]: World Wide Web - Web searching and information discovery - Personalization, World Wide Web - Web mining - Data extraction and integration, World Wide Web - Web mining - Web log analysis

Keywords
user model, short-term user behaviour, website data mining, behavioural changes, session end intent, multi-layer modelling

*Recommended by thesis supervisor: Prof. Maria Bielikova
Defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on [To be specified later]

1. Introduction
User modelling is nowadays a highly important concept in the context of Web. As a process covering acquisition, processing and maintenance of a data about the user, his behaviour and preferences, it has a crucial influence on the task applied: a personalization, recommendation or various predictions on the site. Moreover, indirectly it increases the user experience within a website. Traditionally, the process of user modelling aims on long-term user activity and is focused mainly on the visited content. For this reason, there are modelled mainly preferences and long-term interests of the user.

Traditional modelling approaches capture quality information about the modelled user, they, however, need a sufficient amount of data about the user's activity. For new users, they often suffer by a cold start problem. This is a problem mainly in domains with dynamically changing content and/or minority of loyal users, who visit the site regularly. Additionally, traditional modelling approaches capture mostly the stable long-term user behaviour and thus the updates and new directions in user preferences appear only slowly [4].

Modelling of user preferences offers also information only for selected types of tasks as mentioned personalised recommendation etc. To be able to capture additional traits about the user and his browsing habits in a website, there is need to look for new approaches and also types of a data [2].

In our thesis, we address a task of modelling changes in user behaviour. We focus primarily on the short-term changes, based on which we observe how the user behaviour evolve within a single session. By a session we understand a set of site visits (e.g., page views), realised by the same user, in which couples of consequent page visit time-stamps are distant no more than 30 minutes [8]. Identification of behavioural changes between individual session actions is important to understand actual user behaviour. In this way we are able to react to user intent in an online time and to improve his experience almost intermediately.

2. Modelling User Short-Term Behaviour
User behaviour typically subjects to various factors. On the one hand, there exist long-term interests and knowledge, which are stable and evolve in time only slowly. On the other hand, there are also several important short-

term factors as user's actual context, aim and intent, but also global factors as popular trends. These factors often change in time very quickly and thus together look like a random process.

Modelling of short-term behaviour brings valuable information about recent user activity, but without a support of long-term information, it is mostly insufficient. The short-term behaviour is highly noisy and thus often does not offer information of sufficient quality. According to Xiang et al. [11], for this reason, it is suitable to consider both types of behavioural data together and to use strong sides of them both. An idea of multiple-term modelling later improved Zhou et al. [12], who added third layer of medium-term preferences. In this way, these authors tried to ease differences between modelled information from both layers.

There exist various ways of multiple-term modelling. The first way is based on time windows, where the behaviour is divided into several uniformly sized time windows (e.g. sliding, landmark) or variously sized windows (e.g., damped, adaptive) [1]. Another possibility is a usage of forgetting curve, the process simulating principle of human mind. In this case, the short-term preferences are kept in a memory only for a while (short-term mind) and if not refreshed by a new stimuli meantime, they are forgotten. In the case, that some preference is remembered for a sufficient time, it is moved into long-term memory, where is remembered for a longer time, also in the case of long absence of new stimuli (simulating long-term mind).

3. Behavioural Changes Modelling

User model is typically used as a source of input attributes for a selected machine learning, recommendation or personalization task. According to this task, there is needed to model the data of specific types, for various time periods etc. Kramar et al. [6] recognises following types of the data to be modelled:

- User preferences, interests, goals or attitudes,
- Proficiencies, knowledge,
- Interaction history (user's interaction with the system, tasks performed),
- Stereotypes (e.g., predefined categories).

These types allow detail description of user qualities and descriptive characteristics, they, however, do not allow to observe the updates of individual characteristics in time, their changes or mutual comparisons. These changes are, however, very important for behavioural changes modelling. This kind of data is important in processes such as user behaviour evolution observation, user loss prediction or user leave prediction. In our thesis, we focus on this kind of behaviour, as in this area exist several open research problems.

The task of user loss is typically researched from the long-term perspective, where the user is observed for a longer time period. Based on the activity changes on the level of several days or even weeks, there is possible for example to predict the user attrition. This problem is usually researched as a customer loss [10], student dropout [9] or generally churn rate prediction [7].

In our research, we are interested in the same task applied for a shorter time period. Our aim is to observe behavioural changes within actions of a single session. In this way, we address the actual trend of high domain dynamism. The ability to react to behavioural changes in an online time has a potential to increase the user experience. Typical examples of such reactions are retaining the user within the site, recommending an interesting content to see before leave (e-learning, news) or offering of a discount for user to purchase shopping basket before leave (e-shop).

To capture short-term behavioural changes, we proposed an innovative user model, composed from vector of descriptive attributes and a set of vectors of comparative attributes. The aim of the model is to capture the user behaviour from multiple perspectives, describe it by a high number of attributes and according to specific application to select the important ones. In this way, it is possible to identify significant characteristics in noisy short-term behaviour influenced by various external factors.

Descriptive attributes are used to description of actual session. Here we focus on modelling behavioural attributes capturing data usage, then site characteristics capturing content and structure of pages visited within the session. Detailed description of the individual attributes used can be found in our thesis introducing the model [5].

Comparative attributes are also based on the actual session, in this case, however, in comparison to previous sessions. In this way, we observe how similar is the actual session to previous, or in the other words, if it reminds moments when previous sessions changed (e.g., user ended the session and leaved). Actual session is compared to average of the previous sessions for the last hour, day (short-term behaviour), week, month (long-term) and also for actual part of the day (morning | afternoon | evening) and week (weekday | weekend). In this way, we compare the behaviour in the actual session to the stable long-term as well as long-term behaviour considering actual context or trends. For each time period there is created a individual vector of comparative attributes [5].

To address the cold start problem, which is widespread in dynamic domains with many occasional users, we expand the set of comparative attribute vectors. For each mentioned time layer (hour, day, week, etc.) we compared the actual session with previous sessions of modelled user as well as previous sessions of all users. In this way, we are able to model and to compare the behaviour for users with rich browsing activity in a personalised way. For new and occasional users we compare the actual behaviour at least to average behaviour of all users. An overview of the proposed user model is shown in the Figure 1.

4. Evaluation

To evaluate proposed user model, there is needed to select at first a task in which the model and its attributes will be used. In our case we selected a prediction of user's intent to end the session. More specifically, our aim was to predict if the user will leave the session in next action (or defined number of actions or time interval) or not.

Selected task was evaluated on two datasets from domains with different characteristics, stable e-learning system with lot of loyal users and news portal with majority

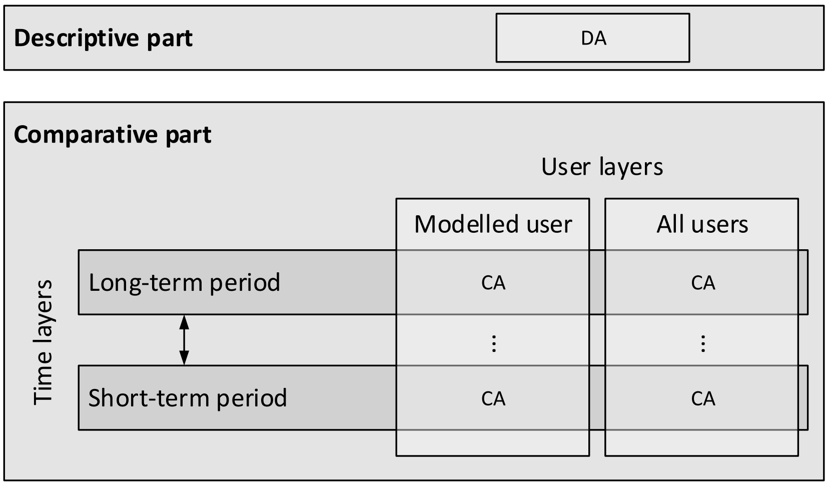

Figure 1: An overview of proposed user model principle. Descriptive attributes part is composed from set of descriptive attributes DA. Comparative attributes part is composed of several sets of comparative attributes CA. Comparative sets are modelled for multiple time layers and user layers. Time layers are captured for last hour, day (short-term period) up to last week, month (long-term period). User layers are captures for modelled user and all site users. The figure is based on [5].

of occasional or one-timer users and dynamically changing articles staying actual only for a while.

4.1 Session End Intent Prediction

We defined the task of session end intent prediction as a binary classification problem over a stream of user browsing activity. For every user action, at first there is predicted if the next action/s will be end of a session or not. After evaluation of the prediction correctness, the observation is used for the classifier training. In this way, the classifier is always up to date, as it is trained continuously [3].

We used polynomial regression classifier, which consider every input variable (model attribute) with a certain importance (weight). The weights are learned by a learning component after every user action processed. This continuous process is ensured by a Stochastic gradient descent algorithm allowing to follow changes in user preferences, habits, seasonality and/or concept drift. An overview of the prediction task is shown in Figure 2.

In addition to prediction of the session end intent in the next action, we explored the prediction within multiple actions in advance and also within defined time interval. The motivation is that an information about the leave in for example two actions (or thirty seconds) may be more useful than information about almost immediate leave in next action. The reason is that a site provider has more time to react and persuade the user to stay longer.

4.2 Prediction Task Challenges

A task of the session end intent prediction brings three main challenges. At first there is high volume of the data to process, secondly there exists a concept drift problem and third, the data are imbalanced.

The high data volume is caused by a fact that users browse websites often very quickly and produce a lot of activity. We reflected this problem by a stream data processing. The second problem is closely related to the first one. As the users browse a lot of pages and the sites dynamically evolve, there often occurs a concept drift. It causes that the subject and intensity of visits changes, so there is needed to adapt quickly. This challenge was handled by a Stochastic gradient descent algorithm.

The third problem lies in data imbalance. Within a session, user typically visits multiple pages, while there exists logically the only one last action. To address this problem, there is needed to balance the data. There exist three techniques - oversampling, undersampling and various importance assigning. To avoid overfitting of rarer class (session end) we refused oversampling, to prevent omitting significant part of data (session continues) we refused also undersampling. Logically as the last option, we decided to assign different importance to the session end/continue observations in the process of attributes weights learning.

4.3 User Model Evaluation Results

The proposed user model was evaluated in two domains (e-learning and news) and three prediction task variants (prediction of session end in next action, multiple actions in advance and defined time interval in advance).

At first, we evaluated basic prediction of end in next action. As the prediction on such a short-term level is difficult because the short-term behaviour is very noisy, we reached for both domains precision at most equal to 60%. For this reason, we further experimented with additional two task variants [3].

We found out, that consideration of multiple actions as a session end increased the prediction quality significantly. We experimented with one, two and three actions in advance. An rapid increase was observed even with consideration of only one action in advance.

A task variant predicting session end within a defined time interval did not performed as well and the precision increase was lower. The reason is that it is more difficult to predict how much time will a user spend within individual pages, even when the user model covers also the characteristics of pages' content. For this reason, we

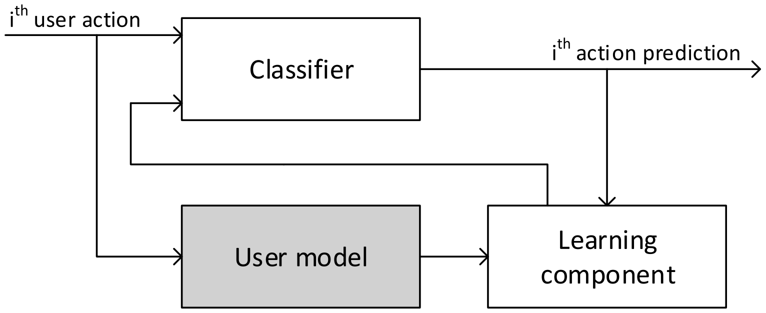

Figure 2: An overview of used prediction method - the user actions are processed continuously as a data stream. They are used as an input for a polynomial classifier. Based on classification result, a prediction is realised. After prediction, its result is used for learning attributes importance (classifier weights). Together with user model attributes, the weights are used as an input for next classifications. The figure is based on our previous paper[5].

further experimented only with a task variant prediction session end within multiple actions.

A reason of high increase of prediction precision when considering multiple actions in advance in comparison to precision of the only last action is that the changes in user behaviour are on the level of session actions very slight. There is very difficult to identify exact changes in user behaviour. The changes exist, it is however almost unable to identify them exactly. Luckily, there is not need to, because as we mentioned earlier, the information about session end within multiple actions has higher practical value (more time to react) than end in next action.

We found out that prediction precision within a one action in advance reached, according to domain, up to 72%, prediction of two actions in advance up to 79% and prediction of three action in advance up to 84% [5]. The results were higher for e-learning domain with longer sessions and more stable loyal users, but news domain results were comparable (lower by several percent at most). The news domain, however contained mostly occasional users and short sessions. The results thus prove the domain independence of proposed user model and its robustness[5].

5. Conclusions

In our thesis we introduced the task of modelling changes in user short-term behaviour. We proposed the model able to capture changes in user behaviour on the level of actions within a session. The main model idea lies in the description of actual user session and its comparison to previous sessions, realised by various users within multiple time periods. In this way, the user model contains a comparison to stable long-term behaviour as well as actual short-term one. Comparison to previous session of modelled user ensures personalised modelling. Potential lack of user's browsing history is compensated by comparison to average behaviour of other users.

In addition, the model is proposed to be domain and language independent, as it uses only generally available usage data describing user activity and indirect descriptors of the structure and content of visited pages. For this reason the model is applicable to almost any site without a need to change its way of user activity logging or content transformation.

To evaluate characteristics of proposed user model, we chose the session end intent prediction task. This task depends mostly on short-term behavioural changes, as it aims at predicting if the user will end his session in the next step or not. By experimenting with a data from two domains with highly different characteristics, we were able to prove that the model is able to capture changes in the user behaviour in general.

The results for both domains showed that such a short-term behaviour is very difficult to recognise exactly. As a rule, the user behaviour before session end however changes. For this reason, the prediction of multiple actions in advance brought a significant improvement in the prediction precision. The high improvement was observed even when considering the only one action in advance. The reason is that based on proposed user model, there is possible to predict that the session end is near.

The slightly better results were reached for e-learning domain with more loyal and stable users who typically visit multiple pages per session. The reason is that in this case, the model data was fully used, the global part (attributes based on comparisons to sessions of all users) as well as the personal part (attributes based on comparisons to sessions of modelled user). In news domain, for high number of users, the personal model part did not contain enough data, so the prediction relied more on global part. Despite that, the prediction based on our proposed user model was able to find and learn important valuable attributes.

The user model was evaluated only by a one task, however, we believe that its potential is higher. It could be used for identification of session author (by looking for the most similar sessions in the past) or fraud detection (by identification of highly unusual user activity in comparison to user's history). Another area for future improvements lies in usage of additional data sources allowing to model new types of data describing different behavioural traits. These could be data from eye trackers, which become nowadays very popular, or another biometric data describing user movements, or gestures.

Acknowledgements. This work was partially supported by grants VG 1/0646/15 Adaptation of access to information and knowledge artifacts based on interaction and collaboration within web environment, KEGA 028STU-4/2017 Innovative methods of teaching informatics in large groups with support for online education and APVV-15-0508 Human Information Behavior in the Digital Space.

References

[1] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy. *Data Stream Mining*, pages 759–787. Springer US, 2010.

[2] O. Kassak, M. Kompan, and M. Bielikova. Personalized hybrid recommendation for group of users: Top-n multimedia recommender. *Inf. Process. Manage.*, 52(3):459–477, May 2016.

[3] O. Kassak, M. Kompan, and M. Bielikova. Student behavior in a web-based educational system: Exit intent prediction. *Engineering Applications of Artificial Intelligence: Mining the Humanities: Technologies and Applications*, 51:136–149, 2016.

[4] M. Kompan and M. Bielikova. Context-based satisfaction modelling for personalized recommendations. In *2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 33–38, 2013.

[5] M. Kompan, O. Kassak, and M. Bielikova. The prediction of session end intent: The short-term user model. *Journal of Universal Computer Science*, page 24, 2017.

[6] T. Kramar, M. Barla, and M. Bielikova. Personalizing search using socially enhanced interest model, built from the stream of user's activity. *J. Web Eng.*, 12(1-2):65–92, Feb. 2013.

[7] F. Li, J. Lei, Y. Tian, S. Punyapatthanakul, and Y. J. Wang. Model selection strategy for customer attrition risk prediction in retail banking. In *Proceedings of the Ninth Australasian Data Mining Conference - Volume 121*, AusDM '11, pages 119–124. Australian Computer Society, Inc., 2011.

[8] C. Liu, R. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 379–386. ACM, October 2010.

[9] M. Tan and P. Shao. Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning*, 10(1):11–17, 2015.

[10] P. Wojewnik, B. Kaminski, M. Zawisza, and M. Antosiewicz. *Social-Network Influence on Telecommunication Customer Attrition*, pages 64–73. Springer Berlin Heidelberg, 2011.

[11] L. Xiang, Q. Yuan, S. Zhao, L. Chen, X. Zhang, Q. Yang, and J. Sun. Temporal recommendation on graphs via long- and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 723–732. ACM, 2010.

[12] B. Zhou, B. Zhang, Y. Liu, and K. Xing. User model evolution algorithm: Forgetting and reenergizing user preference. In *2011 IEEE International Conference on Internet of Things (iThings) & 4th IEEE International Conference on Cyber, Physical and Social Computing (CPSCom), Dalian, China, October 19-22, 2011*, pages 444–447, 2011.

Selected Papers by the Author

O. Kassak, M. Kompan, M. Bielikova. Students' Behavior in a Web-Based Educational System: Exit Intent Prediction. *Engineering Applications of Artificial Intelligence Journal, Mining the Humanities: Technologies and Applications*, volume 51, pages 136–149. Elsevier, 2016.

O. Kassak, M. Kompan, M. Bielikova. Personalized Hybrid Recommendation for Group of Users: Top-N Multimedia Recommender. *Information Processing and Management*, volume 52(3), pages 459–477. Elsevier, 2016.

O. Kassak, M. Kompan, M. Bielikova. User Preference Modeling by Global and Individual Weights for Recommendation. *Acta Polytechnica Hungarica*, volume 12(8), pages 27–41. IEEE, 2015.

O. Kassak, M. Kompan, M. Bielikova. Modelling User Short term Behaviour on the Web: A Survey. *New Generation Computing*, pages 35. Springer, 2017. [Submitted]

M. Kompan, O. Kassak, M. Bielikova. The Prediction of Session End Intent: The Short-term User Model. *Journal of Universal Computer Science*, pages 24. Springer, 2017. [Submitted]

T. Chovanak, O. Kassak, M. Kompan, M. Bielikova. Fast Streaming Behavioural Patterns Mining. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 11. IEEE, 2017. [Submitted]

M. Kompan, O. Kassak, M. Bielikova. Beyond User Preferences: The Short-Term Behaviour Modelling. In Proceedings of the *Temporal Reasoning in Recommender Systems Workshop (RecTemp) collocated with RecSys'17*, pages 3. Ceur, 2017.

T. Chovanak, O. Kassak, M. Bielikova. Behavioral Patterns Mining for Online Time Personalization. *ACM User Modelling, Adaptation and Personalization conference, (UMAP'17)*, pages 2. ACM, 2017.

O. Kassak, M. Kompan, M. Bielikova. Short-term User Behaviour Changes Modelling. *New Trends in Databases and Information Systems: Short Papers and Workshops, BigDap, DCSA, DC*, pages 223–228. Springer, 2016.

O. Kassak, M. Kompan, M. Bielikova. Web User Short-term Behavior Prediction: The Attrition Rate in User Sessions *ECML PKDD 2015 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - Doctoral consortium*, pages 115–124. Library of Aalto University, 2015.

J. Visnovsky, O. Kassak, M. Kompan, M. Bielikova. The cold-start problem: minimal user's activity estimation. Proceedings of the *1st workshop on recommender systems for television and online video (RecSysTV'14)*, pages 4, 2014.