Prediction of User Behavior in a Web Application of the Bank

Peter TRUCHAN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
truchan@swips.sk

In this work, we propose possibilities in measurement and prediction of user behavior and evaluation of measurable users' characteristic metrics. We chose machine-learning algorithms that suited our needs and with the help of these algorithms, we designed model for prediction of user behavior. Input data contain actions and activities made by user in a web application of the bank. We measured more than 130 000 users in the period of three years. We enriched measured data with data from internal database, which contain information about sex and age of the registered users. We used singular value decomposition for dimensionality reduction. Then we used clustering and sequence algorithms to build prediction model. The main contribution of this paper is the proposal of method for building segments of customers that can work with large data in various domains. The only requirement is to build matrix with users' characteristics, visited pages, and sequence of their actions.

In the field of data analysis, there were identified seven different levels of maturity. They are based on the subject's attitude to the data it gathers and the use of this data. The first one is to use only raw data, the second are cleaned data, third are standard, regular reports, the fourth are ad hoc reports and OLAP, the fifth is generic predictive analytics and finally the sixth and seventh are predictive modelling and optimization analysis [1]. First five levels of maturity are usually implemented in all bigger companies. The other levels are not used at all, or only in a limited way. We think that the main reason for that is lack of knowledge, research and established practice in this field. Our aim is to make research in this area and to make predictive modelling better and more accessible. In the field of the predictive modelling there exists a lot of scientific work, but only few of them are related to web data in this form. Server logs and browser extensions are much more widely used than analytical scripts. On the other hand, for the real use in the companies, the server logs are unusable, because of the security concerns.

* Supervisor: Mária Bieliková, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 29–30.

Browser extensions are suited only for limited amount of test users. The only option for company is to use analytical scripts.

The scientific papers with the subject of web user behavior analysis and prediction usually predict the next steps that user is going to take in a application or churn prediction of the user [2]. Usually they use only one artificial intelligence method which is domain depended and not reusable for other website or applications [3]. We gathered data about more than 130 000 users in the period of three years, in the application of internet banking and info site of the bank. Our model is suited for data with little information value – the data we usually get from the website usage. In general, it is the information about visited pages by user, about their time spent, their action, and some long-term characteristic of user as browser, operation system and so on.

Our model is especially suited for prediction of the effect of the change on the visitors. The goal of the model is to answer questions how some change will affect visitors of the webpage. The principle is that we try to find affected segments of users and we predict how they will behave, if we make changes on the webpage. Another usage is to find segments of users, who do not behave according to expectations and then try to predict, what is the best way, and how to change their behavior (this is the part of the optimization analysis). Inputs are data about website usage and user's characteristics and outputs are segments of users and their typical next actions. If we choose only one segment and we know that one part of the segment behave same, we suspect, that if we remove option for them to behave this way (or add option to behave as the other users in the same segment), they will behave as the other users in the same segment.

High level concept is as follows: 1. data gathering by analytical scripts, 2. Data cleaning and preparing for machine learning algorithms, 3. Dimensionality reduction using Singular Value Decomposition, 4. Hierarchical clustering using BIRCH, 5. Representation of sequence of actions, 6. Final data manipulation and visualization

We have proposed a method, which has potential to improve the state of predictive modeling for website owners and web companies. It is based on innovative ways of cleaning and segmenting visitors of web application. We built this model with new approach to combination of machine learning. The results are promising and we expect that we will be able to use it in real world application.

Extended version was published in Proc. of the 12th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2016), STU Bratislava.

References

[1] Chamoni, P. and P. Gluchowski, Integration trends in business intelligence systems: an empirical study based on the business intelligence maturity model. Wirtschaftsinformatik, 2004. 46(2): p. 119-128. Elkan, Ch., Predictive analytics and data mining, p. 15-160

[2] Au, W., Chan, K., Yao, X. 2003. A novel evolutionary data mining algorithm with applications to churn prediction. Evolutionary Computation, 2003. IEEE.

[3] Rahm, E., Hh D., Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 2000.