A Model of Machine Learning Based Interactive E-business Website

Germanas Budnikas

A MODEL OF MACHINE LEARNING BASED INTERACTIVE E-BUSINESS WEBSITE

Germanas Budnikas

Faculty of Economics and Informatics in Vilnius, University of Bialystok, Lithuania
german.budnik@uwb.edu.pl

Faculty of Informatics, Kaunas University of Technology, Lithuania

Abstract. Current online businesses usually contain supporting tools for an instant acquiring offered services and goods. Electronic web site solutions that guide a potential user to a successful finalization of browsing activities, i.e. to purchase finalization play an important role in trading. The proposed in the paper approach presents a model aiming at an assistance in transaction finalization using browsing activity data and personal information. The model is capable to discover missing personal data and to use forecasted values in the research. Results of the applied machine learning techniques are compared while using the whole set of the collected browsing activity data and the data with a reduced number of dimensions.

Keywords: Support Vector Machines, K-means, Artificial neural network, Principal Component Analysis, online customer behavior tracking.

1 Introduction

Practically all of nowadays businesses rely on web sites and web services. Their interaction structure with a visitor can be represented as a two-phase process. During the first phase a user gets some information about a service, during the second phase the user may finalize a transaction with a web site. A transaction finalization is a web site content depended process – it might be a service ordering, commenting, liking something in fb, etc. It is extremely important for business owners to know how web site guests behave online and is it possible to influence their actions. This paper topic addresses these issues and presents results of the research.

Analysis and understanding of web user behavior is a key topic of a behavioral targeting. Behavioral targeting is an evolving area of a web mining that deals with optimization of web online ads based on an analysis of web user behaviors. The research presented in the paper has some similarities to works in the considered field of the study. Methods of behavioral analysis investigate web surfing data gathered mainly from log files. The topic is actively investigated; examples of similar works could be papers by [1], [2], [3] .

adfa, p. 1, 2011.

Approach by [3] suggests a method for monitoring user's online behavior. The method is implemented based on data pulled from log files where HTTP/GET requests are saved when a user clicks a hyperlink. These data are gathered using agent devices installed on user computers. The approach uses Open Directory Project [4] for a categorization of visited web sites. The research emphasizes a creation of behavior profiles with respect to web page visitation event, frequencies and probability distributions, and causality relations or time-dependencies.

Technique by [2] describes the problem of predicting behavior of web users based on real historical data. The data are gathered from user cookie files. An analysis is performed using a statistical decision theory.

Paper by [1] presents a method for modelling and analysis of user behavior in online communities that include personal profiles, wiki, blogs, file sharing, and a forum. The approach implements behavior modelling, role mining and role inference and is based on a statistical clustering.

The approach proposed in the current paper differs from the works listed above by its application area – it operates at Internet level, while [1] and [3] approaches operate at Intranet level. The approach proposed is similar to [1] because they both use a dynamical update of estimations with respect to new data.

The given paper is a continuation of research started in [5] where e-business website data on visitor behavior were collected and used for creation of the model for forecasting online activities of a new visitor. The offered paper differs from the previous one by a wider range of machine learning techniques applied for model creation as well as by taking into a consideration new attributes like user profile data and page or product category visited by a user that are being estimated in case of it absence. Another difference is as follows – since the overall number of tracked attributes had increased, a dimension reduction technique has been employed in order to use the most important components in model analysis. A scoring of the developed model that uses two types of data – original and reduced ones has been prepared in order to evaluate a usability of Principal Component Analysis technique in classification task aiming to discover whether a visitor is willing to finalize a transaction.

The paper is structured in the following way. The second section presents a general architecture of a web site used in the approach. The third section presents a sketch of a procedure of statistical data collecting from a web site. Model creation steps along with the results of machine learning experiments are given in the fourth section. The fifth section briefly outlines a website interaction process based on the proposed machine learning model. Conclusion summarizes an accomplished research.

2 Web site architecture considered in the research

Surfing on web sites usually differs with respect to types of these sites. Open Directory Project (ODP) differentiates the following web site types: Arts, Business, Computers, and 13 more instances. These types generalize manually selected web sites in different languages and are used in various kind of research including the suggested in this paper.

Classification of web sites into types helps in understanding of possible kinds of behavior. Specification of sub-types and its instances is crucial for understanding behavior cases. The paper considers an instance of the Consumer Goods and Services subtype of a business type with respect to ODP classification. Each browsing activity on web sites, especially on business sites, can be logically divided into two parts – introductory that usually includes list of services, descriptions, etc., and (transaction) finalization that could be expressed by paying for services, commenting, fb likes and so on. According to Figure 1, an introductory logical part of a browsing activity may consist of Product Category Selection, Product Selection, Product Related Information Viewing, Delivery and Company Information Viewing; while Check-out and Payment browsing activity corresponds the logical part – transaction finalization.

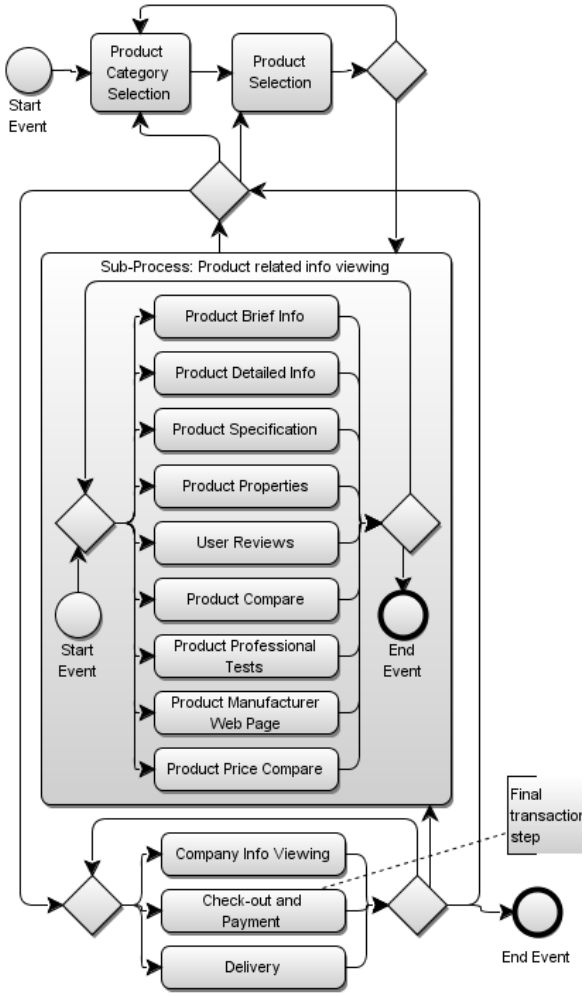

Fig. 1. A generalized view of a user behavior on "Consumer Goods and Services" sites using Business Process Modelling Notation (Source: [5])

Specification of web site surfing activity makes possible to understand a visitor online behavior that can be monitored by using various techniques, e.g., Google Analytics [6] tracking function.

The following 7-tuple could be used for such a specification to define a data-set that is used to monitor customer behavior on an analyzed web site:

<*e*, *y*, *g*, *a*, *u*, *t*, *m*>, where

e is a user browsing *session* during which web site pages are visited;

y is a *category* of a product, viewed by a user. As e-commerce web site may contain a huge number of products (even of the same category), products are differentiated only if they belong to the different categories;

g, a, u correspond to a user gender, age interval and identifier correspondingly set by a cookie file – a tiny text file that contains user visiting information. In case of a new user or not logging to own account, user specific information can be entered in a pop-up window;

t is a kind of an activity or a *task* performed by a user on the web site page like "Product Category Selection", "Viewing Product Price Comparison" (see Fig. 1);

m is activity *t* start time *moment* that is used to differentiate between different browsing sessions.

3 Online visitor behavior tracking

To understand new web site visitor online action, it is needed to track an actual behavior of online visitors. It can be caught using numerous techniques like Open Web Analytics [7] or Google Analytics Event Tracking [8]. Such techniques enable recording user interaction with website elements, such as web page, embedded AJAX page element, page gadgets, and Flash-driven element and so on. Additionally to tracking function, a cookie file is used for unique user identification [9].

Tab. 1. Illustration of tracked data fragment read off from a web site (Source: self-made)

Record number	Session	Product category	Gender	Age interval	User ID	Brief Info	Detailed info	Specification	Properties	User Reviews	Compare	Professional Tests	Manufacturer Web page	Price Compare	Company Info	Delivery	Check-out and Payment
	e	*y*	*g*	*a*	*u*	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_f
R_1	1	1	0	1	1	0	0	0	1	0	0	0	0	1	0	1	0
R_2	2	1	0	1	1	0	0	0	1	0	0	0	0	1	0	1	1

Data about actual on-site surfing is gathered and saved in a form like this

$$\langle e, y, g, a, u, t_1, \dots, t_{n-1}, t_f \rangle,$$

where t_f corresponds the final task. For example, record R_1 (see table 1) represents situation that a user during the first session has visited Product Properties (t_4), Product Price Compare (t_9) and Delivery (t_{11}) web pages and has not finalized the transaction – Check-out and Payment task (t_f) has not been accomplished. Task designations have the following meanings: 0 means a web page has not been visited (i.e., a task has not been accomplished) and 1 means that a web page has been visited. Customer gender information is encoded as 0 for females and 1 for males and 2 for other, age interval has only two intervals – up to 50 and over 50 (surely, that could be expanded onto more detailed parts). User next session (see record R_2 of Tab. 1) consists of visits to the same pages that are marked by grey background color in the table and which ended with a finalization of a transaction. Please note two conflicting records in Tab. 1. They are handled by removing entries that are not ending with transaction finalization while entries with an opposite outcome and the same premises are left. A detailed and formalized description of removing conflicting issues can be found in [5]. An abstract web site, which browsing activity diagram is presented in Fig. 1, was used for an illustration of the proposed technique and up to 3000 visitor behavior data records were used.

4 Model creation steps

Model construction steps can be divided onto 4 parts as depicted on Fig. 2. At the initial step, visitor activities on a web site are tracked and recorded. Website surfers may be logged or not that means visitor profile data – gender and age information may not be accessible. That later fact requires model developing for such data forecasting based on logged in surfers with filled in personal data.

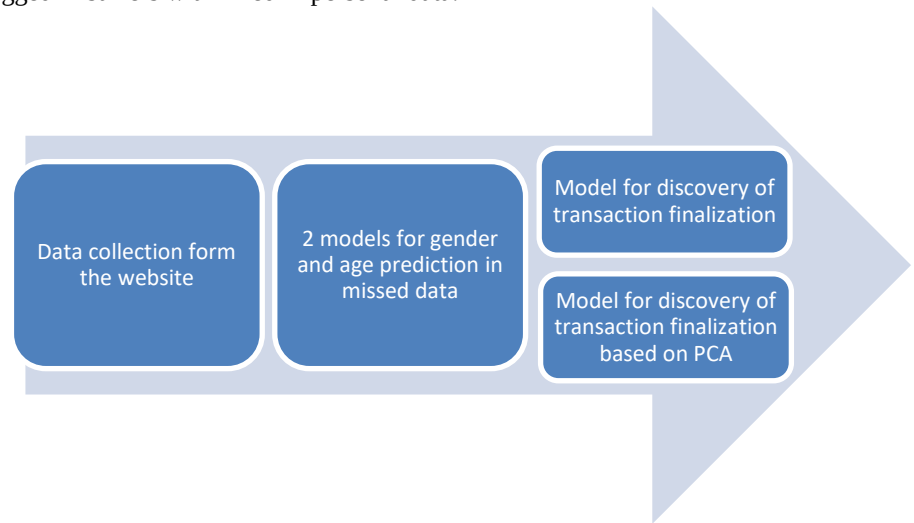

Fig. 2. Model creation steps (Source: self-made)

The first step additionally covers data preparation and processing activities. They include removal of redundant and ambivalent records. An algorithm for processing such

records has been given in [5]. During that phase data for the models to be explained further are being prepared.

As there exists a correlation between visitor age, gender and their browsing patterns ([10], [11], it is important to discover such an information in case of missing data. That could be implemented by building supervised machine learning model based on web surfing activities of registered users.

Binary classification methods were employed for easiness of implementation of female, male and *other* gender discovery. Target classes female/none, male/none were used. A winner class is chosen based on the higher prediction value except the both classes scored greater than 65% (value set experimentally) – an *other* gender is chosen then. The last approach is also applicable in situations when e.g. mother searches for her son's products. The following table summarizes 10-fold cross validation experiments on a discovery of visitor gender in case of anonymous user surfing using three machine learning techniques – artificial neural networks [12], k-means [13], [14] and support vector machines [15].

Tab. 2. 10-fold cross validation results for gender discovery (female, male, other) (Source: self-made)

Artificial neural network	K-means	Support Vector Machines
0.9112	0.8796	0.8612

A classification technique with a highest score obtained based on a given training data set should be employed in the model. In the similar way, two age intervals could be derived. The following table depicts 10-fold cross validation results.

Tab. 3. 10-fold cross validation results for age interval discovery (up to 50, 50+) (Source: self-made)

Artificial neural network	K-means	Support Vector Machines
0.6275	0.6288	0.5530

At the fourth stage, a model for discovering whether a given transaction will be finalized was created. The approach for the model creation at that stage was twofold. The model used all the data taken form the first step (see Fig. 2), namely – number of session, visited types of pages including product categories and personal data (real ones either discovered using aforementioned classification methods). All these data (16 in total) were used as inputs for classification methods. The following table depicts results of cross validation of the applied classification techniques.

Tab. 4. 10-fold cross validation results for transaction finalization discovery (Source: self-made)

Artificial neural network	K-means	Support Vector Machines
0.9761	0.9402	0.8126

Next, architecture and performance characteristics concerning the best scored classification method are presented.

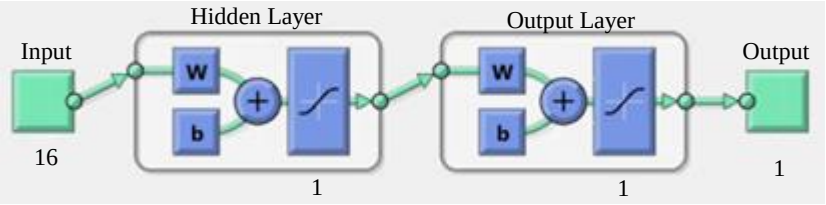

Fig. 3. Artificial neural network architecture used to discover whether a web transaction will be finalized (Source: self-made using Matlab)

Tab. 5. Performance characteristics of artificial neural network used to discover whether a web transaction will be finalized (Source: self-made using Matlab)

	Samples	**MSE**	**%E**
Training	2392	2.61098e-2	2.50836e-0
Validation	149	9.58289e-3	6.71140e-1
Testing	448	2.27893e-2	2.23214e-0

Another competitive approach was applied in order to evaluate how dimension reduction affects classification accuracy. For that purpose, Principal Component Analysis [16] technique was applied. 16 principal components were obtained and the first three of them are depicted on the figure below.

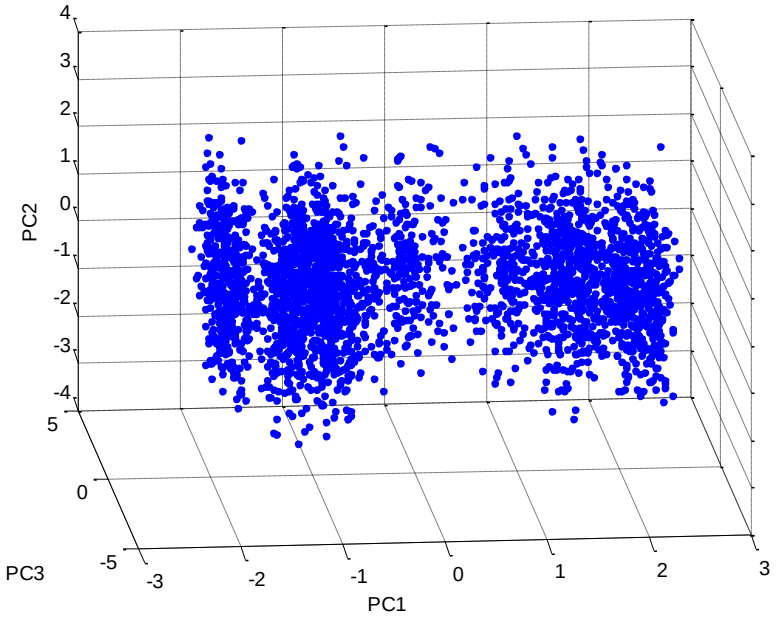

Fig. 4. First three principal components derived from data on website visitor browsing activities using PCA technique (Source: self-made using Matlab)

Below in Tab.6, variance values are given for samples of web surfing data to which Principal Components were applied.

Tab. 6. Variances for all 16 principal components (Source: self-made using Matlab `princomp` function from the Statistics Toolbox)

PC#	Variance
PC1	2.0906
PC2	1.0905
PC3	1.0851
PC4	1.0718
PC5	1.0461
PC6	1.0239
PC7	0.9980
PC8	0.9886
PC9	0.9819
PC10	0.9694
PC11	0.9533
PC12	0.9494
PC13	0.9080
PC14	0.8862
PC15	0.8215
PC16	0.1357

A number of experiments was conducted seeking to discover an appropriate number of principal components to be used as inputs for classification models. The following table presents the best results of the research on dimension reduction while solving classification task.

Tab. 7. 10-fold cross validation results for discovering an appropriate number of principal components (Source: self-made)

PC#	Artificial neural network	K-means	Support Vector Machines
10	0.3868	0.8329	0.8658
11	0.4020	0.8734	0.8883
12	0.4857	0.9012	0.9016
13	0.4943	0.9092	0.8695

The table below (see Tab.8) summarizes results of application of selected machine learning techniques to two groups of tracked web site visitor data: in original form and to the data with reduced number of dimensions using PCA technique.

Tab. 8. Estimation of transaction finalization based on original data and data with reduced number of dimensions (Source: self-made)

	Artificial neural network	**K-means**	**Support Vector Machines**
Original data	**0.9761**	**0.9402**	0.8126
Data with dimension reduction (12 principal components)	0.4857	0.9012	**0.9016**

5 E-business web site interaction with a visitor

A purpose of web site interaction is in providing an assistance for a visitor aiming at influence their willingness to finalize a transaction. Such an interaction has been activated after the visitor had browsed some predefined number of pages or page sections. Based on the constructed models a personal information has been forecasted (see Tab. 9 and Tab. 10) and used in later forecasting tasks defining a willingness of a visitor concerning their final decision.

Tab. 9. Gender prediction for an anonymous visitor based on a partial page visits (7 of 9) (Source: self-made)

Gender	**Artificial neural network**	**K-means**	**Support Vector Machines**
Female	0.6447	0.5982	0.6109
Male	0.5232	0.5146	0.4932
		Resulting predicted gender:	Female

Tab. 10. Age interval prediction for an anonymous visitor based on a partial page visits (7 of 9) (Source: self-made)

Age interval	**Artificial neural network**	**K-means**	**Support Vector Machines**
up to 50	0.5167	0.4973	0.5014
50+	0.4692	0.4618	0.4734
		Resulting predicted age interval:	up to 50

Additionally, the same model has evaluated an influence of each unvisited web page to successful transaction finalization using the machine learning technique that earned best score – artificial neural network without dimension reduction (see Tab.11).

Tab. 11. Prediction of impact levels on the finalization of transaction by anonymous user with predicted personal information for unvisited web pages (Source: self-made)

Unvisited web page	**Artificial neural network**	**K-means**	**Support Vector Machines**
User reviews	0.7517	0.7347	0.7453
Compare	0.6383	0.6228	0.6364
Professional tests	0.7894	0.7629	0.7712

A page (namely – Professional tests), which had been characterized with the highest impact on the finalization, has been advised to visit.

Conclusion and Discussion

The paper describes a model of machine learning based interactive e-business website. The model is illustrated by an example. The model comprises of a usage of three classification techniques that estimate a visitor age interval and a gender, which in their turn used in forecasting the web page with the highest impact on transaction finalization along with actual behavior data. The three machine learning techniques – artificial neural networks, k-means and support vector machines were used to increase credibility of results.

Dimension reduction technique – principal component analysis was used to investigate a possibility to increase a performance of classification techniques. This was reasonable only for support vector machines – an increase of prediction performance was by 8.9%. However, prediction scoring of the three machine learning techniques that used visitor behavior data with the reduced number of dimensions was less than the performance by techniques that used the original learning data.

References

1. Angeletou S., M. Rowe and H. Alani, "Modelling and Analysis of User Behaviour in Online Communities.," in The Semantic Web – ISWC 2011, 2011.
2. Dembczyński K., Kotłowski W., Sydow M., "Effective Prediction of Web User Behaviour with User-Level Models," Journal Fundamenta Informaticae, 89(2-3), pp. 189-206, 2009.
3. Robinson D.J., Berk V.H., Cybenko G.V., "Online Behavioural Analysis and Modeling Meth-odology (OBAMM)," Social Computing, Behavioural Modeling, and Prediction, pp. 100-109, 2008.
4. Xian, X., F. Chen and J. Wang, “An Insight into Campus Network User Behavior Analysis Decision System.,” Taichung, 2014.
5. Budnikas G., "Computerised recommendations on e-transaction finalisation by means of machine learning," Statistics in Transition new series, vol. 16, issue 2, pp. 309-322, 2015.
6. Clifton B. Advanced Web Metrics with Google Analytics, 3rd Edition ed., Indianapolis: John Wiley & Sons, 2012.
7. Jarvinen J., H. Karjaluoto, “The use of Web analytics for digital marketing performance measurement.,” Industrial Marketing Management, vol. 50:, pp. 117-127, 2015.
8. Weber, J. Practical Google Analytics and Google Tag Manager for Developers, Apress, 2015.
9. Aldekhail, M. "Application and Significance of Web Usage Mining in the 21st Century: A Literature Review.," International Journal of Computer Theory and Engineering, vol. 8, pp. 41-47, 2016.
10. Richard M.O., Jean-Charles Chebat, Zhiyong Yang, Sanjay Putrevu, "A proposed model of online consumer behavior: Assessing the role of gender," Journal of Business Research, vol. 63, no. 9–10, pp. 926-934, 2010.
11. Thanuskodi S., "Gender Differences in Internet Usage among College Students: A Comparative Study," Library Philosophy and Practice (e-journal), vol. Paper 1052, 2013.
12. Russell S. and P. Norvig, Artificial Intelligence: International Version: A Modern Approach, 3 ed., Pearson, 2010.
13. Lloyd S. P. "Least squares quantization in PCM, Technical Report RR-5497," Bell Lab, 1957.
14. MacQueen J. B. "Some methods for classification and analysis of multivariate observations.," in L. M. Le Cam & J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297), California: University of California Press, 1967.

15. Christmann I. S. Andreas. Support Vector Machines., Springer New York, 2008.
16. Jolliffe I. Principal Component Analysis., Springer Series in Statistics, 2002.