Web usage mining for adapted and personalized educational space

Daniela OROZOVA[1], Nadezhda ANGELOVA[1], Iva OROZOVA[2]

[1] Department of Computer Science and Mathematics, Trakia University, Stara Zagora, Bulgaria

[2] Ludwig Maximilian University, Munich, Germany

daniela.orozova@trakia-uni.bg, nadezhda.angelova@trakia-uni.bg, orozova96@gmail.com

Abstract: *The paper presents an approach to extract and analyze web data by examining content and learner activities in the learning space. The methodology can be effectively utilized to gain deeper insights into the behaviors of various user groups and the individualized effects experienced during learning in contemporary online educational environments.*

Keywords: Web Metrics, Data Analytics, Web Usage Mining.

1. Introduction

Modern educational spaces are informational and social spaces that integrate heterogeneous learning technologies and different pedagogical approaches. They are a delivery medium for learning materials and educational services for various target groups. The data accumulated by the work of the educational space is constantly increasing and the interest in their use is increasing. Siemens and Long (Long & Siemens, 2014) define Data analytics in Education as "the measurement, collection, analysis and reporting of data about learners and their contexts in order to understand and optimize learning and the environment in which it occurs". By applying Web usage mining in the learning environment, the subject of analysis is the behavior of learners on collections of web pages or resources accessed by groups of learners.

The hypothesis of the current study is that a better understanding of the behavior of various user groups can facilitate the adoption of diverse learning approaches within contemporary educational web spaces.

Conducted research aims to discover patterns in learner behavior and related data collected or generated as a result of their interactions with one or more web sites associated to the learning environment. Such models can be used to understand the behavior of different user groups, to improve the organization and structure of web resources, and to create personalized learning environments by providing dynamic suggestions for resources and activities.

https://doi.org/10.58503/icvl-v19y202403

2. Web analytics in the educational space for adaptive and personalized learning

In modern adaptive learning systems, personalization refers to adapting them to the characteristics of individual learners so that they receive an improved learning process (Brusilovsky, 1998). According to (Bray & McClaskey, 2014), there are three basic approaches to personalized learning: individualization, differentiation, and personalization:

- Individualized training is aimed at learners in need of specific support. Educational goals are usually the same for learners, but they have an individual learning path and their own pace, according to their capabilities and progress in acquiring knowledge. Different ways of teaching and support can be applied, including technological means, oriented to the specific needs and requirements of each learner (Bray & McClaskey, 2014).

- Differentiated (group) learning is based on assigning students to groups based on their levels of knowledge, interest in the field or other characteristics. Teaching methods can be modified to suit the characteristics or preferences of the relevant groups of learners - different approaches, learning resources and tasks can be applied, as well as ways of presenting them.

- Personalized learning is focused on learners who can choose the most suitable way of learning for them. The learning process corresponds most fully to the needs, level of knowledge, interests and individual preferences of different learners. The individual profile of the learner and the individual characteristics that manifest themselves in the learning process are taken into account - preferred ways of presenting learning resources, time planning, choice of learning methods and ways of checking the level of knowledge. Here, the teacher is the student's assistant and mentor and helps to develop his potential and abilities.

The personalization process can be carried out according to the individual needs of given learners (individualized learning), or the characteristics and interests of a group of learners (differentiated learning), as well as according to personal goals (personalized learning).

In order to personalize training, it is necessary to know the capabilities and behavior of the users in order to be able to adapt the different services and resources to them and to improve the interaction. The creation of a model of the user with the characteristics of the learner is the basis of the development of user-oriented systems and of learning systems (Sosnovsky & Dicheva, 2010). On the basis of the data from the model, up-to-date training strategies can be proposed to engage learners more effectively in targeted learning activities.

The creation of models of learners in an e-learning environment aims to represent the acquired knowledge, cognitive skills and interests of the learners as components of a formalized description. Using the Moodog tool (Zhang et al., 2007) installed to the learning environment, statistics are obtained for learners and

for materials that have not yet been used. The Statoodle (Moreno-Marcos et al., 2013) tool uses exported data from Moodle, providing reports and analysis of learner activity, including multiple indicators and log analyzes to detect suspicious test activity due to use of unauthorized materials. From the logs, an Excel file is generated with indicators that contain a summary of each learner's metrics such as: % assignments submitted, % consultations, forum interactions, survey interactions, resource visualizations, % exams taken and failed, etc. In addition, the authors of this paper propose an extension of the learner model, in the context of activities in a web environment, as a basis for personalizing the learning process in the web educational space.

An approach to learner modeling is presented, based on an analysis of their behavior in the learning environment and the web space. The model includes basic elements with data about the learners in the learning environment: grades received on individual assessed units and data from the electronic platform such as time spent in the e-learning course, downloaded materials, submission of independent works, course assignments, projects, etc. (Popchev & Orozova, 2019). A new element in the model is an assessment of the content and the extent of used websites related to the field of study. For this purpose, a web analysis of the educational space is carried out.

Web analytics pertains to the systematic measurement, collection, analysis, and reporting of data from the Internet, aimed at better understanding of user interactions with websites. This process encompasses a range of web metrics as delineated in Google Analytics (Clifton, 2008).

- *page views* - the number of views of a web page accessible by a human visitor (without robots);
- *visitors* - the number of unique visitors to a website;
- *pages/visits* - the number of pages that visitors saw while visiting the site;
- *time on site* - length of time spent by all visitors to the website;
- *stickiness* - the ability of the web page to keep a visitor;
- *frequency* - the number of visits made by a visitor to the site (loyalty indicator);
- *recency* - number of days that have passed since the last visit of a given visitor to the website;
- *length of visit* - visit time spent by visitors to the website (in seconds);
- *depth of visit* - the number of pages visited by a visitor during one visit to the website and a number of other metrics.

Web analytics uses a variety of techniques to evaluate website traffic, including server-side and client-side data collection. The methods of collecting data from the server are related to extracting and analyzing data mainly from log files. When collecting data from client sites, data about a page visitor is sent to a tracking server, usually via a JavaScript code (or tag) inserted into the HTML page.

A visitor's actions can be tracked and additional information collected (Zumstein & Kaufmann, 2009). Google, WebTrends, Nedstat and other companies provide web analytics software using page markup. Google Analytics is the most used free program (Clifton, 2008).

In the context of analyzing learner activity, data concerning customers, including their visit history and behavioral patterns, as well as user profiles, is of significant interest. Data from customer relationship management (CRM) systems are also stored in operational databases. To provide a holistic view of learners in an online environment, data from various sources must be collected, cleaned and transformed before being integrated and analyzed in data warehouses.

3. Stages in the Web usage mining process

The subject of analysis is learner behavior on collections of web pages or resources accessed by groups of learners in a learning environment. Extracted data comes from: web server log files, site content, visitor data collected from external channels, additional learning environment data.

In the analysis of the web space, the main data presents the visitor sessions and the actions undertaken from the point of entry until exit from the website. However, several challenges arise in acquiring accurate data, including the use of proxy servers and anonymizers, dynamic IP addresses, missing references due to caching, and the limitations of servers in differentiating between distinct visits without supplementary tools, etc. (Bing, 2011).

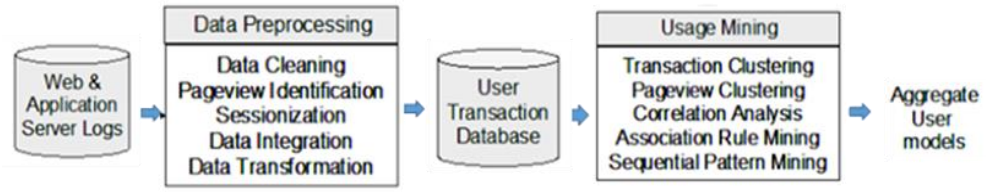

Figure 1. Web usage mining process

The quality of the discovered patterns depends on the quality of the data on which the algorithms are applied. Data preprocessing involves several stages (Figure 1):

In the *data cleaning stage*, the following actions are performed: deletion of unrelated fields in the server logs files, removal of wrong references, addition of missing references due to caching (after session), etc.

In the *data integration stage*: synchronization of data from multiple server logs, integration of metadata, integration of demographic or log data, etc.

Data transformation is related to: user identification, data is associated with relevant sessions (sessionization), data reduction may be required (dimensionality reduction, ignoring certain page views / page elements), etc.

Session reconstruction is a fundamental procedure that involves accurately distinguishing the activities associated with various individuals and attributing multiple visits to the same individual. Different methods are used for user identification: built-in session IDs; user registration and login to the site; use of cookies that record the ID of the client machine; software agents loaded on the browser that send user data, etc.

After preprocessing the collected data, a set ***P*** of visited web pages is formed and a set of user transactions ***T*** - sequences of page views that can be associated with weights are obtained:

- number of visited pages: $P=\{p_1, \ldots, p_n\}$;
- set of user transactions: $T=\{t_1, \ldots, t_m\}$, where each transaction t_i contains a subset of P;
- each transaction $< (p_1^t, w(p_1^t)), (p_2^t, w(p_2^t)), \ldots, (p_l^t, w(p_l^t)) >$

is an ordered sequence of pageviews of length *l*, where each *w* corresponds to a weight defining the importance of the displayed page. These weights can correspond to explicit user ratings or to the transactions collected on the network - this can be the duration of the page visit in the session.

In cases where the order of activities in a transaction does not matter to the modeling algorithm (such as clustering, extracting associative rules, etc.), the transaction can be represented as a vector of length *n*: where the weight is 0 if the corresponding page is not present in *t* otherwise corresponds to the meaning of the page in *t* (Bing, 2011).

Using a matrix of user views of the web resources (0 - not visited or 1 - visited) we have, for example, the data for four learners and five observed resources:

Table 1. Matrix of user views of web resources

Learners/Resources	**Resource** A	**Resource** B	**Resource** C	**Resource** D	**Resource** E
Learner 1	1	1	0	0	0
Learner 2	0	0	1	0	0
Learner 3	1	0	0	0	1
Learner 4	1	0	0	1	1

Information about the content of the resources can be integrated such as: text features from the web content represent the underlying semantics of the pages. So, for example, if we put the text features: "association analysis", "regression analysis", "cluster analysis" and "classification analysis", we can associate them with the content of web resources as:

Table 2. Matrix of web resources

Resources /Topics	Association analysis	Regression analysis	Cluster analysis	Classification analysis
Resource A	0	1	1	0
Resource B	1	0	0	1
Resource C	1	1	0	0
Resource D	0	0	1	0
Resource E	0	0	0	1

By performing the multiplication of the Learners/Resources x Resources/Topics matrices, the user visits matrix of web resources can be transformed into a content matrix (Bing, 2011):

Table 3. Content matrix

Learners /Topics	Association analysis	Regression analysis	Cluster analysis	Classification analysis
Learner 1	1	1	1	1
Learner 2	1	1	0	0
Learner 3	0	1	1	1
Learner 4	0	1	2	1

After the data extraction, cleaning, integration and storage stages, algorithms can be applied to extract knowledge from the data and create models. Different data analysis techniques can be used such as: clustering the rows of the matrix can reveal learners with common interests, associative rules can be extracted from the data or outliers can be detected.

4. Data modeling for Web usage

Various types of analysis can be successfully applied to the received data, such as: session analysis, statistical processing; OLAP and Data Mining techniques (Ranieri & Silvestri, 2007).

Session analytics is a commonly used form of server data analysis and is related to viewing individual or group sessions. A general idea of the typical behavior of learners in the learning environment is obtained. In addition, specific environmental problems can also be identified. Analytics is about processing a large amount of data. A very commonly applied form of data analysis is creating summaries (grouping) of data by predefined units (eg days or sessions). Advantages of this approach are that it gives a quick overview of how a site is being used and requires minimal memory and processing power to process the data. But no deeper analysis of the data is provided.

Online Analytical Processing (OLAP) - tools that are used for multidimensional data analysis and allow changes in the level of aggregation for multiple dimensions. Dimensions are indices by which the individual values of the

multidimensional array are indexed. The number of dimensions is not limited and thus a hyper-cube of data is obtained. Each dimension has its own attributes (characteristics). The main operations of OLAP are: aggregation (summarization of data depending on the hierarchy levels in a given dimension), drill-down (outputs detailed data on the given dimension), filtering (outputs data that meets a certain condition formulated on one dimension), generation of section, selection of a subset (scoping), rotation (pivot), etc. Typically, these tools connect to Data Warehouse. This is a very flexible approach to data analysis, but requires significant resources.

Data Mining is the process of searching for hidden data and regularities, previously unknown and practically useful, necessary for making decisions. The emphasis is not only on extracting new facts, but also generating testable hypotheses. Knowing the basic algorithms and data analysis approaches such as: classification analysis, regression analysis, association analysis, cluster analysis, noise analysis, etc., on which the various Data Mining techniques are based, can help extract useful knowledge from the accumulated web space data.

In the *classification task*, a finite set of objects is set, for which it is known which classes they belong to, and the class affiliation of the remaining objects is unknown. An algorithm must be constructed that classifies an arbitrary object by specifying the value of the target attribute. When the possible values of the target attribute are only two, then we have a "binary" classification problem, in the other case - "multiclass". In general, the algorithm works by generating a series of random rules. In the prediction task, the data is previously divided into two groups with mutually exclusive elements - training and test data sets. Based on the first set, a model of the data is built by generating rules and selecting those that best fit the data. The process is repeated a certain number of times until a rule that satisfies (almost 100%) the training data is found. The rules are then verified against the test data and the model is evaluated.

Regression analysis provides an answer to the question of what are the factors that contribute to a particular outcome. It shows the interrelationships between quantities that can be interpreted as causal. This is a statistical analysis designed to give a quantitative expression of the effects of a given group of variables X_1, X_2, ..., X_p, which are conditionally called "independent" on another variable Y - "dependent". The main idea is to search for a natural functional relationship of the form: $Y = f(X_1, X_2, \ldots, X_p)$. The equation is a regression equation. The objectives of regression analysis are to determine how and to what extent the dependent variable varies or changes as a function of changes in the independent variable. The variables whose variation we want to explain or predict are the effects. The independent variable is the cause.

Associative analysis studies the frequency of co-occurrence of facts. This analysis is concerned with discovering “associative rules” specifying conditions for attribute values that occur frequently together in a given data set. Associative rules

have the form: X ≥ Y. Thus, the rule consists of two parts: X is the conditional (antecedent) part, and Y is the logical consequence (resultant part) (Marr, 2015).

In *cluster analysis* the goal is to group n number of objects into k number of groups, called clusters, using p number of features (variables). Thus, the cluster is formed by similar objects, regardless of their classes. The goal is to reveal a possibly hidden grouping of the objects. An important division of clustering procedures is depending on whether the number of clusters is specified in advance. A large variety of procedures arises from the rules used to create the clusters (Hornick et al., 2006). The most frequently used methods of this type are: "nearest neighbor", "furthest neighbor", "on centroids", etc.

Noise analysis - the data may contain objects that do not support the underlying behavior or pattern of the data. They are called deviations (Outliers) and are defined as noise. Sometimes, however, these data are more interesting than other cases. Deviations are the differences between measured values and corresponding expectations based on previous or normative values. A data analysis task is, when a set of deviations is detected, to create a description of the characteristics of the deviations, explain the reason for it, propose an action to bring the values back to their expectations, etc.

5. Data mining in the web educational space

Applying Data mining techniques in the web educational space can facilitate the resolution of numerous tasks, including the identification of intriguing interrelationships, the classification of learners into distinct groups, the enhancement of learning resource planning, the assessment of interest levels in various resources, among other applications.

Traditionally, in learning environments, content is delivered to learners based on predefined rules such as: "if the learner has clicked on resource A and their current review score is more than 4.5" then "add a link to resource C". Similar rules generate learner recommendations based on the responses and ratings of other "similar" learners.

By applying associative analysis to data, associative rules can be discovered and patterns of page sequences can be extracted from pages in certain sessions, creating opportunities to predict likely preferred pages to access and proactively offer new learning resources. Example of association link between page views:

"x% of the time A appears in a transaction, B appears after it in z number of transactions."

Through Cluster Analysis of web page visit data, important visitor segments or links in using resources in the same sequence can be discovered. Groups of learners with common interests can be revealed, based on the content available or by other means. The resulting groups can be subjected to Classification Analysis,

taking into account other characteristics such as current success, time spent in the learning space, etc. and to create profiles of certain types of learners.

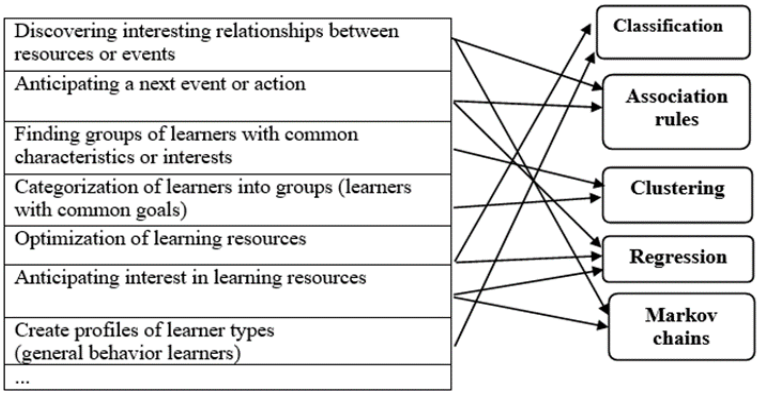

Figure 2. Applications of Data mining tools in an educational environment

On the other hand, combined information about the structure and use of web resources can be used to analyze and optimize the learning environment.

The analysis of web resources associated with the training course provides a foundation for making informed decisions regarding the enhancement of these resources. It allows for the identification of challenges faced by learners in locating relevant topics and facilitates the incorporation of additional literature, including new websites that pertain to the most frequently searched subjects. This methodology enables the modeling of learners within the educational context and permits the monitoring of their online activities, thereby accumulating valuable data regarding their engagement in the learning environment.

Alongside with the benefits of the suggested methodology there are some constraints regarding the quantity of web resources that are monitored and analyzed, as well as the number of students involved.

6. Conclusion

The proposed approach promotes the creation of an innovative learning environment and is the next step in the digitization of education. New machine learning technologies and the technical power of modern computers create the real opportunity for innovative applications based on the integration of heterogeneous data. The choice of algorithm when solving a given problem depends on many factors, for example, the size and type of data, also its quality and quantity, depends on what the result will be used for, as well as the time required to reach a result (Popchev & Orozova, 2023). Cloud technologies combined with big data management, storage and processing technologies such as Hadoop/HDFS, Spark, Hive, provide access to shared data, analytical tools and opportunities to use massive computing resources.

Acknowledgment

This work is supported by the frames of Bulgarian National Recovery and Resilience Plan, Component "Innovative Bulgaria", the Project № BG-RRP-2.004-0006-C02 “Development of research and innovation at Trakia University in service of health and sustainable well-being”.

REFERENCES

Bing, L. (2011) *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg, 2nd ed., Chapter written by Bamshad Mobasher, ISBN 978-3-642-19459-7.

Bray, B. & McClaskey, K. (2014) *Make learning personal: The what, who, wow, where, and why*. Corwin Press. Thousand Oaks, CA.

Brusilovsky, P. (1998) Adaptive educational systems on the world-wide-web: A review of available technologies. In *Proceedings of Workshop" WWW-Based Tutoring" at 4th International Conference on Intelligent Tutoring Systems (ITS'98), San Antonio, TX*.

Clifton, B. (2008) *Advanced Web Metrics with Google Analytics*. Wiley, New York, USA.

Hornick, M., Marcade, E. & Venkayala, S. (2006) *Java data mining: strategy, standard, and practice*. *A practical Guide for Architecture, Design, and Implementation*. Morgan Kaufmann, 1st ed.

Long, P. & Siemens, G. (2014) Penetrating the fog: analytics in learning and education. *Italian Journal of Educational Technology*. 22(3), 132-137.

Marr, B. (2015) *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*. John Wiley & Sons Ltd.

Moreno-Marcos, P. M., Barredo, J., Muñoz-Merino, P.J. & Delgado Kloos, C. (2023) Statoodle: A learning analytics tool to analyze moodle students’ actions and prevent cheating. *Lecture Notes in Computer Science*. Springer Nature, Switzerland, 736–741.

Popchev, I. & Orozova, D. (2023) Algorithms for Machine Learning with Orange System. *International Journal of Online and Biomedical Engineering (iJOE)*. 19(04), 109 -123, doi: 10.3991/ijoe.v19i04.36897.

Popchev, I. & Orozova, D. (2019) Towards Big Data Analytics in the e-learning Space. *Cybernetics and Information Technologies*. 19(3), 16-25. doi: 10.2478/cait-2019-0023.

Ranieri, B. & Silvestri, F. (2007) Dynamic personalization of web sites without user intervention. *Commun*. ACM 50(2), 63-67.

Sosnovsky, S., & Dicheva, D. (2010) Ontological technologies for user modelling. *International Journal of Metadata, Semantics and Ontologies*. 5(1), 32-71.

Zhang, H., Almeroth, K., Knight, A., Bulger, M. & Mayer, R. (2007) Moodog: Tracking students’ online learning activities. In: *EdMedia+ Innovate Learning. Association for the Advancement of Computing in Education (AACE).* pp. 4415–4422.

Zumstein, D. & Kaufmann, M. (2009) A Fuzzy Web Analytics Model for Web Mining. *IADIS European Conference on Data Mining*. pp. 59-66. ISBN: 978-972-8924-88-1.