RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS No. 104
Business Informatics 16 2010

Wiesław Pietruszkiewicz[*], Dorota Dżega[**]
[*]West Pomeranian University of Technology, Szczecin, Poland
wpietruszkiewicz@wi.zut.edu.pl
[**]West Pomeranian Business School, Szczecin, Poland
ddzega@zpsb.szczecin.pl

AN APPLICATION OF DATA MINING IN THE MANAGEMENT OF E-LEARNING PLATFORM

Abstract: In this paper we present an appalication of data mining in education, where the management of e-learning platform was supported by extraction of users activity features and clustering of users' profiles. It allowed us to identify groups of users with a similar activity and to observe their performance. While the majority of other researches focus on the analysis of students, we investigated teachers' behaviour. The experiments presented herein were performed on the real data coming from Moodle platform. We have proposed a smoothing model in the form of a dynamic system which was used to transform the logged events into time series of activities. These series were later used to cluster teachers' performance and to divide them into three groups: active, moderate and passive users. We claim that an increase of e-learning quality requires responsible persons not only to observe students but also to evaluate teachers. The gathered information may be used to identify potential teachers problems, e.g., technical difficulties or low e-learning platform skills.

Keywords: clustering, e-learning, kernel k-means, k-medoids.

1. Introduction

The control of users is very important in e-learning. There were performed various researches about users modelling, e.g. Ventura et al. [2008] showed how the decision trees could be used by teacher to find relationships between students marks and their activity. The usage of neural networks to predict students marks was presented in [Delgado et al. 2006], while research showed in [Romero et al. 2008] contained comparison of different classifiers used as students' grades predictors. The other researchers proposed the data mining procedures to analyse the usage of e-learning courses to support their improvement [Blondet Baruque et al. 2007], analysed a data mining application to quality control of e-learning [Balogh 2009] or used case-based reasoning in distance learning [Shen et. al. 2003]. In [Tang, McCalla 2002] clustering was used to analyse learners' behaviour, being treated as a sequences of virtual movements. A similar clustering based analysis of virtual steps was proposed in [Mor,

Minguillón 2004], while the identification of incorrect students behaviour was presented in [Yu, Own, Lin 2001]. Automatic recommendation of relevant materials for students, done by machine learning, was presented in [Markellou et al. 2005] and the analysis of students' learning sequences was introduced in [Pahl, Donnellan 2003]. A detailed survey of educational data mining can be found in [Romero, Ventura 2007].

If we compare teachers to students receiving the grades, we can notice that it will be difficult to form any precise rule that could be used to evaluate instructors' performance. While the lectures in traditional learning can be visited and their frequency is easy to check, the same procedure for e-learning is impossible.

It this paper we present how data mining can be used to observe teachers and to support the management of e-learning platform. A schema of the proposed approach may be found on Figure 1. This process contains five major steps:

– generation of reports with events registered for users,
– conversion of events sets to time series,
– calculation of activity attributes,
– grouping users with a similar activity attributes,
– identification of users with 'active', 'moderate' and 'passive' profiles.

Each step will be explained in the further parts of this article. It must be also noted that we will focus on teachers, but a similar analysis could be done for students. However, due to the lack of literature positions presenting educational data mining applications oriented on teachers, we decided to investigate their behaviour. The data source used herein was a Moodle platform, used to teach 105 courses with 43 teachers.

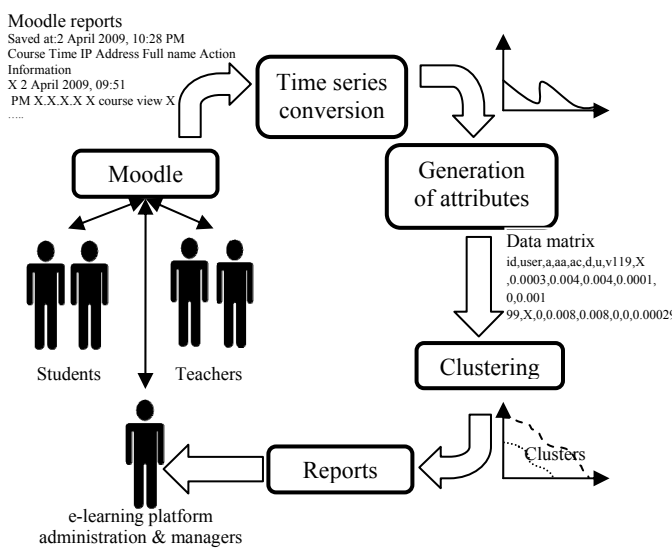


Figure 1. The steps of process of users' activity modelling

It shall not be assumed that observation procedures in e-learning must be used only as a negative stimulation mechanism. We rather think that such capability is required by the whole process of learning, where low activity of instructors could mean, e.g., technical problems with the e-learning platform. Moreover, a common thinking about learning is limited to teachers and students, neglecting management staff, while this third group is responsible for ensuring the quality of the whole process.

2. Activity components

We have used Moodle reports to prepare preliminary files which were later converted to well structured array of attributes, being considered as a representation of users' activity in six areas. All possible events, that were found in these files, were gathered in Table 1.

Table 1. The events logged for all six activity components

Component	Events for component
1	2
'Add'	email add mail, forum add discussion, forum add post, blog add, message add contact, course add mod, assignment add, quiz add, resource add, glossary add entry, label add, glossary add, survey add, chat add, wiki add, lesson add, notes add, workshop add
'All changes'	user login, user logout, email add mail, forum mail blocked, forum add discussion, message write, message history, chat talk, forum add post, course recent, forum user report, chat report, user update, upload upload, user change password, forum subscribe, course report outline, course report stats, course user report, blog add, course report log, course report participation, forum delete post, message add contact, forum search, forum update post, course add mod, assignment add, forum delete discussion, assignment update grades, course delete mod, quiz editquestions, quiz add, course report live, email reply, forum stop tracking, forum start tracking, message remove contact, forum subscribeall, forum unsubscribe, resource add, course update mod, assignment update, resource update, course editsection, glossary delete entry, glossary add entry, course update, chat update, login error, forum move discussion, scorm report, label add, glossary add, glossary update, survey add, quiz report, quiz update, quiz close attempt, quiz continue attempt, chat add, wiki her house, wiki edit, wiki update, wiki info, wiki add, label update, glossary update entry, glossary approve entry, wiki links, lesson update, lesson add, wiki strippages, notes add, calendar edit, message unblock contact, message block contact, course unenrol, workshop update, workshop add, wiki diff, wiki attachments, discussion mark read
'Delete'	forum delete post, forum delete discussion, course delete mod, glossary delete entry
'Update'	user update, forum update post, assignment update grades, course update mod, assignment update, resource update, course update, chat update, glossary update, quiz update, wiki update, label update, glossary update entry, lesson update, workshop update

1	2
'View'	user view, user view all, forum view forum, course view, email view course mails, email view mail, forum view forums, forum view discussion, resource view, chat view, chat view all, glossary view, blog view, notes view, resource view all, scorm view, scorm pre-view, forum view subscribers, scorm view all, quiz view, assignment view submission, assignment view all, assignment view, glossary view all, game view all, game view, survey view form, quiz preview, quiz view all, quiz review, wiki
	Interview your partner about his, wiki view, wiki view all, workshop view, workshop view all
'All actions'	All events for 'Add', 'All changes', 'Delete', 'Update' and 'View'

To convert the reports into the measures of activity, it was necessary to propose a general guideline explaining what these values should denote. The proposed criteria of activity contained two rules:

I. The more various actions teacher takes, the higher value of activity should be.

II. Promote more regular teachers, than those who have logged events grouped only in short sessions.

The later part of article will introduce appropriate equations, delivering values of activity components.

3. Modelling users activity

The reports generated by Moodle, like other web logs, contain only information about time of events, not about users' activity between these events. The naïve approach to this problem would be to calculate an average number of events and use it as an activity measure. However, this approach would not distinguish events spread equally from events cumulated in large groups. To create time series of activity from a set of impulses we have chosen smoothing form of process equation written in form of a dynamic system, which in the space of state variables may be introduced as [Wan, van der Merwe 2001]:

$$x_{k+1} = F(x_k, u_k), \tag{1}$$

$$y_k = H(x_k), \tag{2}$$

where: x_k – vector of state variables,
u_k – vector of input variables (control variables),
y_k – vector of output variables (observed variables).

The structure of this system was shown on Figure 2. Function F presents characteristics of the modelled process, while function H is used to represent possibility of x measurements, as there are some situations that measured y output values are aggregates of the state variables.

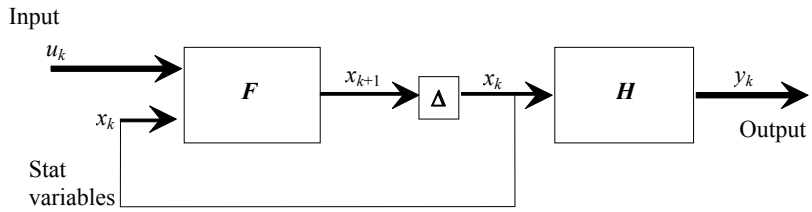


Figure 2. Basic schema of a dynamic system

As the transforming system was a logic construction and all state variables were accessible, we could omit observation equation (2) and focus on process equation (1). However, before we will present the equations used to smooth data, it is necessary to explain the main concept of smoothing. The idea of transformation of events into time series of activity was graphically presented on Figure 3. Each event stimulated activity at the moment it occurred and the activity value was decreasing in next steps. This assumption makes it possible to create a smooth time representation of impulses (logged events).

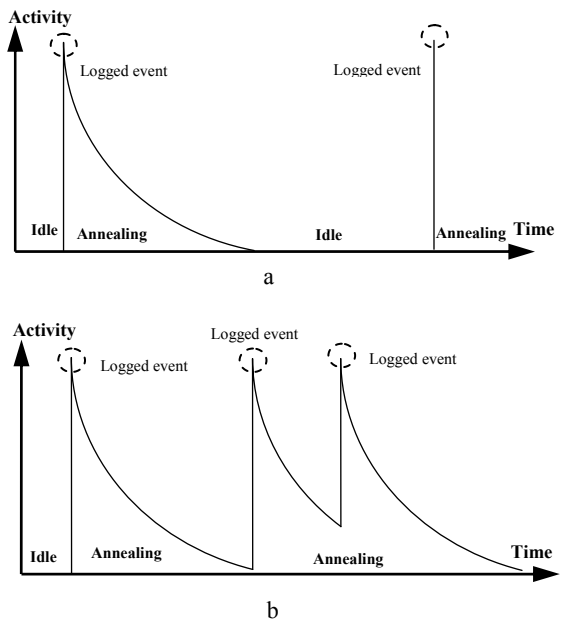


Figure 3. Smoothed time series for: high activity (a) and low activity (b)

As modelled system had to be time-discrete, we have used a time window check if any event occurred during the analysed period. Overall activity was a vector containing six components mentioned in Table 1. Moreover, it was essential to keep

separated different kinds of activity and do not use only ‘all actions’ component, because other five components are important features that distinguish more active users from the others.

To make our algorithm running we have created an appropriate process equation. Denoting activity vector by X, events vector by U and limiting state variables to $x \in$ <0,1>, the process equation was formed as equation (3).[1]

$$X_{n+1} = \min\left(\begin{bmatrix} p & 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 0 & 0 & 0 \\ 0 & 0 & p & 0 & 0 & 0 \\ 0 & 0 & 0 & p & 0 & 0 \\ 0 & 0 & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & p \end{bmatrix} * \begin{bmatrix} x_n^a \\ x_n^a \\ x_n^a \\ x_n^d \\ x_n^u \\ x_n^v \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} u_n^a \\ u_n^a \\ u_n^a \\ u_n^d \\ u_n^u \\ u_n^v \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}\right). \quad (3)$$

As it can be noticed, activity forms peaks and later, due to P-diagonal matrix, descents to zero. Value of multiplier p is a ratio of fading, i.e., the lower value it has, the faster the fading is. The length of time window was 15 minutes. At each step the value of u components was set to 1 if any event occurred for this activity component, otherwise it was set to 0.

In the next step this model was used to calculate time series for all six activity components. An example of these series for one of the users was presented on Figure 4. These time series represented a measurable value of activity (higher values for frequent events, which is a characteristic feature of the active teachers). Due to the time scale of observation, the peaks were squeezed on chart and formed vertical black lines. These time series were transformed for each user to form the vectors containing average values of six components. This step must be explained, as it may seem that it is opposite to Rule II mentioned in Section 2. However, we must remember that the length of time window was set to 15 minutes. No matter how many events occurred during each discrete time window, the value was set to be 1 if period contained any logged events or to be 0 otherwise. The length of window was short enough to catch most of time-distributed events, but was long enough to group events occurred in short sessions. Let us assume we have a situation where two users have equal number of logged events, but one of the users was working frequently, while the other used platform during short sessions. Using the proposed time conversion, most of events will fall into the same time windows for the second user whose activity will be low, while the activity of the first user will be higher. In the result, by averaging the values of activity, it was possible to convert each time series into one value and to obey both rules introduced in Section 2.

[1] We see this equation to be a some sort of parallel to the Page’s method used to detect changes in time series [Page 1954].

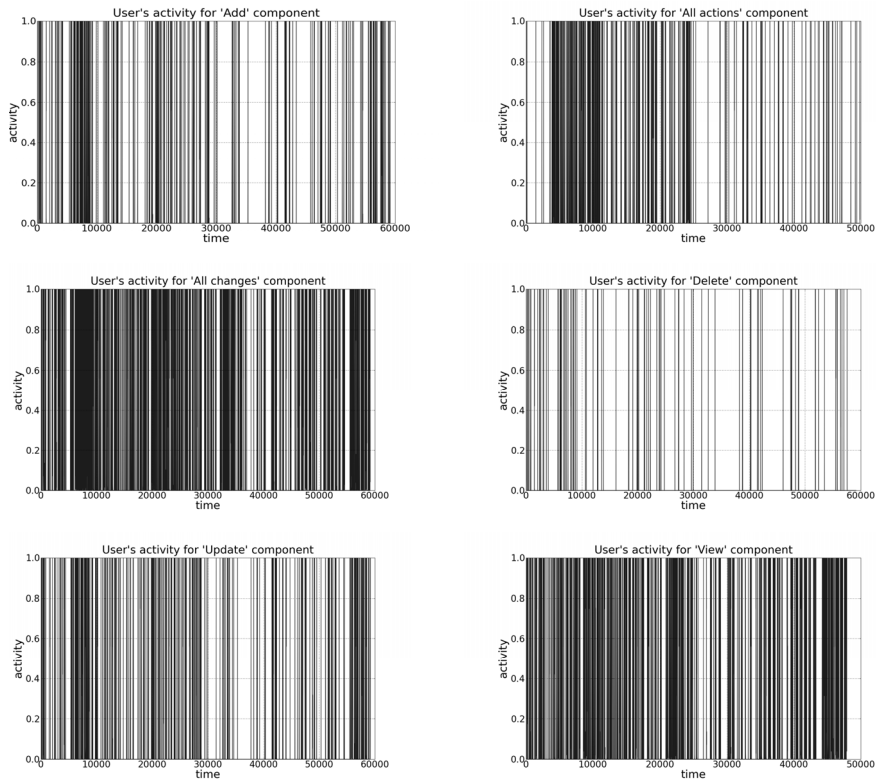


Figure 4. Sample of six components of activity

It is important to remember that courses available on Moodle contained all materials, in SCORM form, necessary for students to learn the course subject. Tasks, tests, games, glossaries, lessons, workshops, quizzes and other activities were additional, thus some teachers were more active engaging students to different tasks, while other were more passive, i.e., responding to students' problems, moderating forums and chatting with students. Thus, we do not consider passive users to be a negative group, but as a group that must be observed more carefully and to eventually support them technically.

As there is no attribute that could be used to evaluate of the quality of model (comparing to, e.g., students' grades) we decided to evaluate the achieved results using experts' opinions. The e-learning management staff analysed the grouped users' profiles and confirmed how the results of clustering were coherent with their observations.

4. Results of clustering

The clustering was performed using two algorithms, i.e., k-medoids [Mierswa et al. 2006] and kernel k-means [Camastra, Verri 2005]. As we had 43 observations for different lecturers, we set the number of clusters to 8 and later joined clusters into 3 groups: 'Active', 'Moderate', and 'Passive' users. The distributions for 'Add' vs. 'View' and 'Add' vs. 'Delete' for both algorithms were presented on Figure 5. The cluster No. 7 was empty for kernel k-means algorithm. The grouped clusters representing activity profiles for both algorithms were presented in Table 2.

Table 2. Clusters for activity profiles

	Kernel k-means	k-medoids
Active	3,6	0,4,5
Moderate	5	2
Passive	0,1,2,4	1,3,6,7

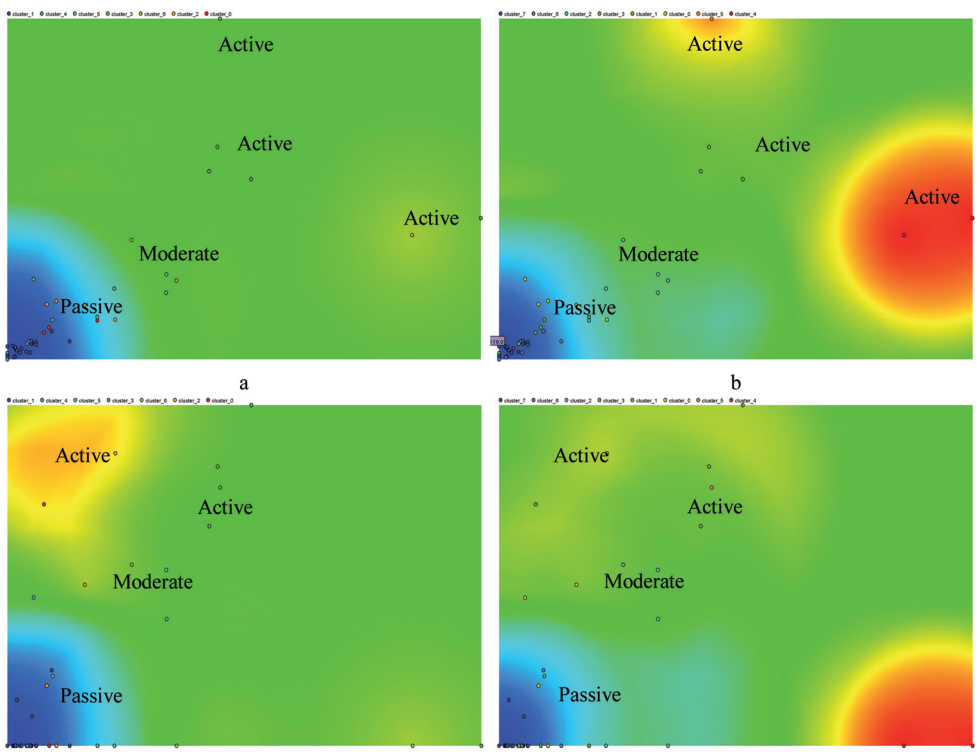
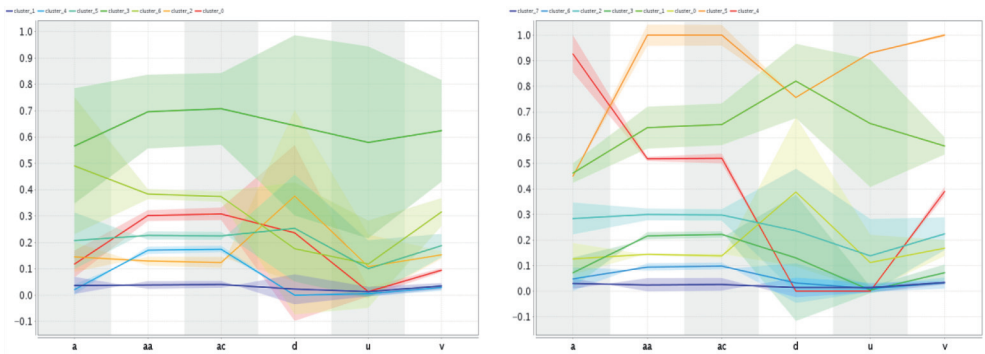


Figure 5. Clusters density in 'Add' vs. 'View' dimensions for kernel k-means (a) and k-medoids (b) algorithms or in 'Add' vs. 'Delete' dimensions for kernel k-means (c) and k-medoids (d) algorithms



a b

Figure 6. Normalised deviation of attributes for kernel k-means (a) and k-medoids (b) clusters

The analysis of deviation (Figure 6) showed that factor 'Delete' was differing much for clusters generated by kernel k-means, while 'Add' was the most useful factor for k-medoids clusters. Comparing both algorithms, we have selected k-medoids as it was more coherent with users' observation done by Moodle administrative staff and the clusters generated by this algorithm were easier separable.

5. Conclusions

In this paper we presented how the hidden features may be extracted from Moodle and be used to support the management of e-learning. The results of clustering described herein were achieved using attributes pre-processed by a dynamic system, smoothing data coming from Moodle reports. We have proposed the activity measures for six areas, that were later used by k-medoids and kernel k-means algorithms to group users with similar activity profiles. While many researchers focus on modelling students behaviour, we presented an analysis of teachers. We also think that future extension of e-learning platforms must support management staff in observing users' behaviour. This process should be done by procedures that would observe users' activity, measure it and supply easily accessible and understandable information for e-learning managers. We see the proposed approach to be the basis of a supporting mechanism, supplying information allowing to early identify teachers problems and to transform the e-learning management to a more effective and pro-active way of. Our future plans involve separations of activity components onto more detailed subsets. Another potential area of further research is a linkage between teachers activity and students results. This will join previous research done for students with research proposed in this paper.

References

Balogh I. (2009), Use of data mining tools in examining and developing the quality of e-learning, [in:] *Proceedings of LOGOS Open Conference on Strengthening the Integration of ICT Research Effort*, Budapest.

Blondet Baruque C., Amaral M.A., Barcellos A., João Carlos da Silva Freitas J.C., Juliano Longo C.J. (2007), Analysing users' access logs in Moodle to improve e learning, [in:] *Proceedings of the 2007 Euro American Conference on Telematics and Information Systems*, Faro.

Camastra F., Verri A. (2005), A novel kernel method for clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 5.

Delgado Calvo-Flores M., Gibaja Galindo E., Pegalajar Jiménez M.C., Pérez Piñeiro O. (2006), Predicting students' marks from Moodle logs using neural network models, [in:] *Current Developments in Technology-Assisted Education*, FORMATEX, Badajoz.

Markellou P., Mousourouli I., Spiros S., Tsakalidis A. (2005), Using semantic web mining technologies for personalized e-learning experiences, [in:] *Proceedings of the web-based education*, Grindelwald.

Mierswa I., Wurst M., Klinkenberg R., Scholz M., Euler T. (2006), Yale: Rapid prototyping for complex data mining tasks, [in:] *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*, Philadelphia.

Mor E., Minguillón J. (2004), E-learning personalization based on itineraries and long-term navigational behaviour, [in:] *WWW Alt. '04: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, New York.

Page E. (1954), Continuous inspection schemes, *Biometrika*, Vol. 41.

Pahl C., Donnellan C. (2003), Data mining technology for the evaluation of web-based teaching and learning systems, [in:] *Proceedings of the congress e-learning*, Montreal.

Romero C., Ventura S. (2007), Educational data mining: A survey from 1995 to 2005, *Expert Systems with Applications*, Vol. 33, No. 1.

Romero C., Ventura S., Espejo P.G., Hervás C. (2008), Data mining algorithms to classify students, [in:] *Proceedings of Educational Data Mining 2008: 1st International Conference on Educational Data Mining*, Québec.

Shen R., Han P., Yang F., Yang Q., Huang J. (2003), Data mining and case-based reasoning for distance learning, *Journal of Distance Education Technologies*, Vol. 3, No. 1.

Tang T.Y., McCalla G. (2002), Student modelling for a web-based learning environment: A data mining approach, [in:] *Eighteenth National Conference on Artificial Intelligence*, American Association for Artificial Intelligence, Menlo Park.

Ventura S., Romero C., Hervás C. (2008), Analyzing rule evaluation measures with educational datasets: A framework to help the teacher, [in:] *Proceedings of Educational Data Mining 2008: 1st International Conference on Educational Data Mining*, Québec.

Wan E.A., van der Merwe R. (2001), The unscented Kalman filter, [in:] *Kalman Filtering and Neural Networks*, John Wiley & Sons, New York.

Yu P., Own C., Lin L. (2001), On learning behaviour analysis of web based interactive environment, [in:] *Proceedings of ICCEE*, Oslo/Bergen.

ZASTOSOWANIE EKSPLORACJI DANYCH W ZARZĄDZANIU PLATFORMĄ E-LEARNINGOWĄ

Streszczenie: od kilku lat obserwujemy intensywny rozwój usług edukacyjnych świadczonych w środowisku Internetu. Na rozwój ten w istotny sposób wpływa rozwój systemów informatycznych zdalnego nauczania oraz rozwój narzędzi wspomagających tworzenie treści kursów e-learningowych. Za rozwojem technologii powinien także podążać rozwój kadry dydaktycznej. Nowoczesna kadra dydaktyczna powinna nie tylko przekazywać wiedzę, ale przede wszystkim aktywnie uczestniczyć w budowaniu ścieżek edukacyjnych. W artykule prezentujemy jeden z mechanizmów wspomagających rozwój e-learningu. Jest nim kontrola aktywności i zachowań nauczycieli przeprowadzona z wykorzystaniem metod eksploracji danych. Artykuł omawia kolejne etapy tego procesu, prowadząc od ekstrakcji danych, poprzez ich konwersję aż do wniosków wyciągniętych z wykonanej klasteryzacji.