INFORMATION
SYSTEMS IN
MANAGEMENT

Information Systems in Management (2014) Vol. 3 (1) 77–88

ON USING DATA MINING TECHNIQUES FOR CONTEXT-AWARE STUDENT GROUPING IN E-LEARNING SYSTEMS

DANUTA ZAKRZEWSKA

Institute of Information Technology, Lodz University of Technology

Performance of an e-learning system depends on an extent to which it is adjusted to student needs. Priorities of the last ones may differ in accordance with the context of use of an e-learning environment. For personalized e-learning system based on student groups, different distribution of the groups should be taken into account. In the paper, using of data mining techniques for building student groups depending on the context of the system use is considered. As the main technique unsupervised classification is examined,. Context parameters depending on courses and student models are tested. Experiment results for real student data are discussed.

Keywords: E-learning, Data Mining, Student Grouping, Context-Awareness

1. Introduction

In e-learning good student performance depends on en extent to which educational environment is tailored to learners' profile [1]. Grouping students of similar characteristics enables to adjust educational system to groups of colleagues, who should learn together from the same resources. However features of educational environment should also differ depending on the context of the system use. Context aware differentiation of teaching materials seems to be an important feature of an e-learning system [2]. The research aims at examining data mining techniques namely unsupervised classification methods for student groups' creating, in different context of the educational system usage.

The paper is organized as follows. In the next section, literature review concerning context-aware personalized e-learning systems, student grouping as well as application of data mining in e-learning will be presented**.** Then, context aware models in e-learning systems will be described. The following section will be devoted to application of clustering techniques for student grouping taking into account context of use.

Finally, the case study of student profiles characterized by learning styles will be considered and experiments on building context aware groups for real students data will be described and discussed. The paper will be finished with concluding remarks and future research presentation.

2. Related work

Data mining techniques were very often applied as personalization tools in e-learning systems. The broad review of the research in that area can be found in [3, 4]. Cluster analysis was used to group students on the basis of their behaviors (see [5, 6]) or individual traits (see [4] for example). Many authors connected clustering with different data mining techniques to increase efficiency of obtained results. Shen et al. [7] applied cluster analysis together with sequential pattern mining to group students according to their learning activities. Tang and McCalla [8], in turn, integrated clustering technique together with collaborative filtering for building recommendations of course contents. The last technique combined with association rules mining was often used for building student recommendations (compare [9, 10]).

In e-learning systems, student groups were often built for recommendation purposes. Authors grouped students taking into account their behaviors, pages they visited or historical navigational paths ([6, 8]), as well as learner cognitive styles or usability preferences [4].

Many of the researchers emphasized an importance of a context in a personalization process (see [2, 11, 12], for example). The broad review of context parameters as well as context aware e-learning systems was presented in [13]. Context-awareness was very often considered for recommendation purposes. Andronico et al. [14] built multi-agent system to suggest students educational materials taking into account learners' behavior and their preferences while using different mobile devices. Rosaci and Sarné, in turn, considered both: student's profile and an exploited device [15]. Their recommendations were built on the basis of the time spent by student on the particular Web site, taking into account type of a device used for navigating. Zaïane [16] proposed an agent, which aims at recommending learning activities or shortcuts in course web-sites, considering learners' historical activities. Using of Naïve-Bayes models for building context-aware group recommendation was proposed by Zakrzewska [17].

3. Context-aware models in e-learning

Dey [18] defined context as “any information that can be used to characterise the situation of an entity”. Context of use plays an important role in e-learning. Students’ needs may differ depending on the situation of use and what is more, different student features may be important in the cases of different courses.

Dey [18] said that “the system is context-aware if it uses context to provide relevant information or services to the user, where relevancy depends on the user’s task”. In educational systems context-awareness should be taken into account in order to obtain its personalization features. In that case, relevant information can have the form of teaching materials tailored into learners’ needs according to the usage situation.

Das et al. [19] distinguished three types of context parameters in e-learning: personal, abstraction and situation. The first one is connected with student personal information, personality type and the level of expertise. Situation context describes learner situation, network and device he uses [19]. In our considerations we will focus on abstraction context which concerns the information of student preferences and learning styles, in the situation of the course, that learner attends. We will assume that the context model is different depending on the course that student enrolls on.

Let us assume that each student *ST* is described by *N* attributes, which may indicate their learning styles or other preferences. A tuple *ST* of a student is of the form:

$$ST = (st_1, st_2, ..., st_N), \quad st_i \in DOM(S_i), \tag{1}$$

where *DOM*(S_i) stands for the domain of S_i .Further, we will assume, that the attributes are of different importance, depending on the context of an educational system usage. Let the context *CN* will be described by *N* weight parameters:

$$CN = (cn_1, cn_2, ..., cn_N), \tag{2}$$

where $cn_i \geq 0$, $i=1,2,...,N$; mean the importance of the *i*-th attribute of the student model in the context *CN*. We will also suppose that

$$\sum_{i=1}^{N} cn_i = 1, \quad cn_i \geq 0, \quad i = 1,..., N. \tag{3}$$

If any of the attributes has no significance in the considered context, respective weight value is equal to 0.

To take into account the context of use, we will include context vector into the grouping process. As the most important, student features according to the biggest weights of the context vector will be considered. Cluster analysis of students’ data was also broadly examined in [4], where different algorithms were considered in order to build groups of students of similar needs. Investigations, presented there, showed advantages of unsupervised classification. The main problem, which arises in the current research, consists in including context into cluster analysis tool.

Assuming that each course can be modeled by a vector of weights related to respective student attributes, we propose to present context-aware grouping as clustering problem with weighted distance function. Then, similarity of group members will be measured by a function, where contribution of each of the attributes depends on the respective weight values.

Let x and y be vectors $x = (x_1,x_2,...,x_N)$ and $y = (y_1,y_2,...,y_N)$ and let $w = (w_1,w_2,...,w_N)$ denote a vector of weights, where w_i is non-negative for every $I = 1,..,N$ and fulfils (3). Then a weighted distance function d_w will take the form:

$$d_w(x,y) = \sum_{i=1}^{N} w_i d(x_i, y_i) \tag{4}$$

Such way of including weights into a distance function will not change none of its metric properties and will enable taking into account priorities of each of the attributes. However according to such approach only distance based clustering can be applied, but the results of the proposed technique should not depend on the choice of an algorithm.

Obtained student groups should be different depending on the context and their quality should be measured by taking into account the most important attributes, from the context point of view. Students from the same group should have the most similar features, which tutors decide as the crucial in the considered context. The cluster quality can be examined by calculating standard deviations separately for each attributes within clusters. The smallest value should concern features of the biggest importance.

4. Learning styles and usability preferences case studies

Let us consider student models based on their dominant learning styles. We will examine the model which was often used to ensure adaptivity features of e-learning systems [20], proposed by Felder and Silverman [21]. It is based on “Index of Learning Style” (ILS) questionnaire developed by Felder and Soloman [22]. The results of ILS questionnaire indicate preferences for 4 dimensions from among excluding pairs: *active* vs. *reflective*, *sensing* vs. *intuitive*, *visual* vs. *verbal*, *sequential* vs. *global*. The index, obtained by each student, has the form of the odd integer from the interval [–11; 11], assigned for all the dimensions.

Thus student learning style model *SL* is represented by 4 integer attributes:

$$SL = (sl_1, sl_2, sl_3, sl_4) = (l_{ar}, l_{si}, l_{vv}, l_{sg}) . \tag{5}$$

Element l_{ar} means scoring for *active* (if it has negative value) or *reflective* (if it is positive) learning style, and respectively l_{si}, l_{vv}, l_{sg} are scores for all the other dimensions, negative values are in cases of *sensing*, *visual* or *sequential* learning styles, while positive values are in cases of *intuitive*, *verbal* or *global* dominant learning styles.

That way $N = 4$. We will consider building student groups in the context of different courses, then each course *CR* will be modeled by a vector of weights, which signify the importance of different learning style preference for the course:

$$CN = CR = (w_{ar}, w_{si}, w_{vv}, w_{sg}) . \tag{6}$$

As the second student model usability preferences will be considered. As the most important design categories, deciding on Web sites usability, which should be evaluated by users, Marsico and Levialdi [23] mentioned information representation and appearance, access, navigation and orientation as well as the informative content architecture of the sites. Investigations presented in [24] showed that students put special attention to graphical attractiveness of Web sites and the efficiency which means a short time of loading the sites. Students also emphasized the importance of advanced search possibilities. Consequently, five preferences for portal features are taking into account: informative contents, graphics, navigation, efficiency and search possibilities. Students were asked to score the importance of each of the feature, assigning from 1 to 5 scores. Values equal to 1 or 2 mean that a student does not put attention to the portal characteristic, 3 means that a learner does not distinguish the importance of considered feature from among the others, finally values 4 or 5 mean that the usability trait is important for the student. Let *SU* denote student usability preference model. Taking into account the meaning of the score values, *SU* is represented by 5 attributes:

$$SU = (su_{1,} su_2, su_3, su_4, su_5) , \tag{7}$$

where su_1 means scoring for importance of informative content, su_2 scoring for importance of graphics, su_3 scoring for importance of navigation, su_4 scoring for importance of efficiency of the system and finally su_5 means scoring for the importance of search possibilities. Then $N = 5$ and consequently, the respectful course model of weights is of the form:

$$CR = (wu_1, wu_2, wu_3, wu_4, wu_5). \tag{8}$$

5. Experiment results and discussion

Experiments aimed at checking, how including context into clustering process will change student group characteristics. Two main attribute categories were considered: dominant learning styles and usability preferences. The research was done on the basis of experiments conducted on the two trial sets of students from Technical University of Lodz: the set A of 22 learners studying the same master course of Information Systems in Management, and the set B of 56 part-time Computer Science students, who were also graduates of other programs. Firstly, students filled ILS questionnaire and answered questions concerning usability needs then groups of similar preferences were built, taking into account each course context. Finally grouping effects were evaluated.

For the grouping purpose 8 different courses were considered. During the courses: *CR1, CR2, CR3, CR4* students were grouped according to their dominant learning styles. For the courses *CR5, CR6, CR7, CR8,* in turn, usability preferences were taken into account. All of them characterised by different weight vectors.

It was decided that a student model for the course *CR1* is global and all the learning styles attributes are equally important, while during the preparation of the course *CR2* only dimensions: *active/reflective* or *visual/verbal* should be taken into account. This dimension has the highest priorities in both of the courses: *CR3* and *CR4*, however its importance is much bigger for the course *CR4*. In this course the dimension *sensing/intuitive* is not considered. Learning style dimensions' weights for all the courses are presented in Table 1.

Table 1. Weights for courses. Case of dominant learning styles

Course	w_{ar}	w_{si}	w_{vv}	w_{sg}
CR1	1/4	1/4	1/4	1/4
CR2	1/2	0	1/2	0
CR3	1/6	1/6	1/2	1/6
CR4	1/20	0	9/10	1/20

Models are global in the case of the course *CR5* and all the usability preferences attributes are of the same priority. Educational materials prepared for the course *CR6* were distinguished depending on three usability preferences: informative content, graphics and navigation, not taking into account efficiency and search possibilities. Informative content is of the highest priority in the course *CR7*, while all the others attributes are of the same importance. For the course *CR8*, navigation is the most important feature, informative content is the second one, all the other features are of the same priorities. Weights of usability preferences for all the courses are presented in Table 2.

During experiments students' data from the sets A and B were clustered, taking into account all the courses' needs. Context of the course was included as weight values into distance function. Students were divided to 5 groups, the number for which clustering schemas was stated to be optimal in many cases, while student grouping [25].

Table 2. Weights for courses. Case of usability preferences

Course	wu_1	wu_2	wu_3	wu_4	wu_5
CR5	1/5	1/5	1/5	1/5	1/5
CR6	1/3	1/3	1/3	0	0
CR7	1/3	1/6	1/6	1/6	1/6
CR8	1/5	1/10	1/2	1/10	1/10

Groups were built by using k-means algorithm implemented in the Open Source Weka software [26], taking into account Manhattan distance function.

During the process of quantitative analysis all the obtained clusters were compared, taking into account course context. Standard deviations within clusters were used to examine its qualities. Mean values of attributes decided of cluster profiles. The results for the set A and courses *CR1*, *CR2* and *CR4* are presented in Table 3 . Clusters obtained for the course *CR3* were of the same parameters as the ones obtained for the course *CR1*. Weights close or equal to 0 changed the structure of obtained groups. '-' means that the attribute was not taken into account during the clustering process.

To examine, how weights included into clustering process influence grouping effects, standard deviations for the whole sets and the most important attributes were calculated and compared: l_{ar} l_{vv} for the set A and courses *CR1, CR2, CR4* and l_{vv} for the set B and courses: *CR1, CR2* and *CR4*.

In the first case averages of standard deviations are respectively equal to 1.92, 1.70 and 3.23 for l_{ar}, and 2.87, 1.32 and 2.32 for l_{vv}. In the case of set B obtained values are respectively: 2.83, 2.15, 4.49 for l_{ar} and 2.7, 2.69 and 1.55 for l_{vv}. It can be easily noticed that removing not important attributes from the clustering process ameliorated the quality of obtained groups.

Table 4 contains respective values of means and standard deviations of attributes within clusters obtained for the set B and courses *CR1, CR2* and *CR4.* Similarly to the previous case the results got for the course *CR3* were the same as for *CR1*.

Table 3. Set A. Mean values of attributes within clusters

Crs	Clst. No	Inst.	Mean values				Standard deviations			
			l_{ar}	l_{si}	l_{vv}	l_{sg}	l_{ar}	l_{si}	l_{vv}	l_{sg}
CR1	1	4	-5	-5	-5	-4	1	1.93	1	4.76
	2	4	-6	-4	-9	-1	2.83	5.74	1	2
	3	4	6	-6	-10	-3	2.83	1.91	2.83	4.43
	4	2	2	-9	1	1	1.41	0	5.66	0
	5	8	-5	1	-3	3	2.07	2.25	3.85	1.41
CR2	1	8	-5	-	-5	-	1.07	-	2.83	-
	2	5	-7	-	-9	-	1.79	-	0	-
	3	5	5	-	-11	-	3.63	-	2.61	-
	4	1	3	-	5	-	0	-	0	-
	5	3	-1	-	-3	-	2	-	1.15	-
CR4	1	2	-4	-	-6	-8	1.41	-	1.41	1.41
	2	8	-6	-	-9	-1	1.49	-	1.69	3.20
	3	2	3	-	-10	-7	2.83	-	1.41	2.83
	4	2	-1	-	7.5	2	5.66	-	3.54	1.41
	5	8	-1	-	-5	1	4.78	-	3.54	1.28

Table 4. Set B. Mean values of attributes within clusters

Crs	Clst. No	Inst.	Mean values				Standard deviations			
			l_{ar}	l_{si}	l_{vv}	l_{sg}	l_{ar}	l_{si}	l_{vv}	l_{sg}
CR1	1	8	3	-11	-5	-5	2.39	2.33	2.83	4
	2	13	-7	-3	-1	-3	2.23	3.38	3.38	2.66
	3	12	-3	-6	-7	-5	2.58	3.13	1.51	3.57
	4	15	7	-7	-3	-1	3.40	2.56	3.64	2.33
	5	8	-9	7	-9	1	3.55	2.71	2.14	3.21
CR2	1	8	1	-	-3	-	1.51	-	2.14	-
	2	13	-7	-	-1	-	2.51	-	3.15	-
	3	4	7	-	-2	-	1.63	-	4.76	-
	4	12	2	-	-5	-	1.97	-	1.80	-
	5	19	-5	-	-9	-	3.12	-	1.61	-
CR4	1	12	1	-	-3	-3	5.52	-	0.98	1.71
	2	6	-5	-	3	-2	3.77	-	2.34	3.88
	3	13	-3	-	-7	-5	3.80	-	1.70	2.24
	4	13	-1	-	-5	-1	4.82	-	1.30	2.09
	5	12	-5	-	-9	1	4.55	-	1.44	2.97

Table 5 contains respective values of means and standard deviations of attributes within clusters obtained for the set A and courses *CR5* and *CR6*. Similar results for the set B are presented in Table 6. For both of the sets results obtained for the course *CR7* and *CR8* are the same as for the course *CR5*.

For the courses *CR5* and *CR6* averages of standard deviations of su_1 ,su_2,su_8, which are the most important in the course *CR6* were considered. In the case of the set A the averages for *CR5* are respectively equal to 0.1, 0.61, 0.29, while for *CR6* they take the value: 0, 0.61, 0.1. Similarly to the case of learning styles preferences one can conclude that removing of attributes of the low importance from the clustering process ameliorated the quality of obtained groups.

Table 5. Set A. Mean values and standard deviations of attributes within clusters

Crs	Cl. No	Inst.	Mean Values					Standard deviations				
			su_1	su_2	su_3	su_4	su_5	su_1	su_2	su_3	su_4	su_5
CR5	1	3	4	4	4	2	3	0	0.58	1	0.58	1.15
	2	4	5	3	4	5	4	0.5	0	0	0.5	0.5
	3	5	4	3	5	4	3	0	0.89	0.45	0.45	0.55
	4	3	5	5	4	4	3	0	0.58	0	0.58	0.58
	5	7	5	4	5	4	3	0	0.98	0	0.69	0.95
CR6	1	4	4	3	4	-	-	0	0.5	0.5	-	-
	2	6	5	5	4	-	-	0	0.98	0	-	-
	3	2	4	3	5	-	-	0	0	0	-	-
	4	7	5	4	5	-	-	0	0.97	0	-	-
	5	3	4	4	5	-	-	0	0.58	0	-	-

Table 6. Set B. Mean values and standard deviations of attributes within clusters

Crs	Cl. No	Inst.	Mean Values					Standard deviations				
			su_1	su_2	su_3	su_4	su_5	su_1	su_2	su_3	su_4	su_5
CR5	1	7	4	3	4	3	2	1	1.13	1.15	1.11	0.53
	2	22	4	4	4	4	4	0.58	0.66	0.70	0.73	0.77
	3	1	1	1	3	5	5	0	0	0	0	0
	4	17	4	3	5	5	4	0.66	0.87	0.24	0.94	0.79
	5	9	5	4	4	3	3	0.33	0.44	1.05	0.97	0.60
CR6	1	8	5	4	3	-	-	0.52	1.19	0.46	-	-
	2	16	4	4	4	-	-	0.65	0.72	0	-	-
	3	1	1	1	3	-	-	0	0	0	-	-
	4	21	4	4	5	-	-	0.54	0.85	0	-	-
	5	10	5	4	5	-	-	0	1.26	0	-	-

In the case of the set B, for the course *CR5* average standard deviations for first three attributes are equal respectively to 0.51, 0.62, 0.42.. For the course *CR6*, in turn, they are equal to 0.34, 0.80 and 0.12. Again, it could be easily noticed that removing of the attributes before the grouping process improved the quality of obtained clusters.

6. Conclusion

In the paper application of unsupervised classification for student groups' creating, in different context of the educational system usage, is investigated.

The context of the courses is considered, taking into account such student features as dominant learning styles and usability preferences. In the proposed method course context is presented as a vector of weights, included into clustering process. The effects are evaluated on the basis of the experiments done on the datasets of real students.

Tests showed that using of weights can ameliorate qualities of obtained groups if they differ significantly. What is more qualities of the biggest clusters were better from the most important attributes point of view.

Future research should consist on further experiments concerning more number of attributes of different meaning and importance as well as using different clustering technique. More precise cluster validation technique should be also considered.

REFERENCES

[1] Beaudoin M.F. (2002) *Learning or Lurking? Tracking the "Invisible" Online Student,* Internet & Higher Educ., 2/2002, 147-155.

[2] Schmidt A., Winterhalter C.(2004) *User Context Aware Delivery of E-learning Material: Approach and Architecture,* J. Univers. Comput. Sci.,10/2004, 38–46.

[3] Romero C., Ventura S. (2010) *Educational Data Mining: a Review of the State of the Art*, IEEE T. Systems, Man & Cybernetics, Part C: Applications & Reviews 6/2010, 601-618.

[4] Zakrzewska D. (2012) *Eksploracja danych w modelowaniu użytkowników edukacyjnych systemów internetowych,* AOW EXIT, Warszawa, Poland.

[5] Perera D., Kay J., Koprinska I., Yacef K., Zaïane O.R. (2009) *Clustering and Sequential Pattern Mining of Online Collaborative Learning Data,* IEEE T. Knowl. Data En., 6/2009, 759-772.

[6] Talavera L., Gaudioso E. (2004) *Mining Student Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces*, Workshop on Artificial Intelligence in CSCL. 16th European Conference on Artificial Intelligence, 17–23.

[7] 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, 3/2003, 46–58.

[8] Tang T., McCalla G. (2005) *Smart Recommendation for an Evolving E-learning System*, International Journal on E-Learning, 1/2005, 105–129.

[9] Minaei-Bidgoli B., Tan P., Punch W. (2004) *Mining Interesting Contrast Rules for a Web-based Educational System*, The Twenty-First International Conference on Machine Learning Applications, 1-8.

[10] Wang F. (2002) *On Using Data-Mining Technology for Browsing Log File Analysis in Asynchronous Learning Environment*, Conference on Educational Multimedia, Hypermedia and Telecommunications, 2005–2006.

[11] Jovanowić J., Gašewić D., Knight C., Richards G. (2007) *Ontologies for Effective Use of Context in E-learning Settings*, Educ. Technol. Soc., 10/2007, 47-59.

[12] Yang S.J.H. (2006) *Context Aware Ubiquitous Learning Environments forPeer-to-Peer Collaborative Learning*, Educ. Technol. Soc., 9/ 2006, 188-201.

[13] Das M.M., Chithralekha T., SivaSathya S. (2010) *Static Context Model for Context Aware E-learning*, International Journal of Engineering Science and Technology, 2/2010, 2337–2346.

[14] Andronico A., Carbonaro A., Casadei G., Colazzo L., Molinari A., Ronchetti M. (2003) *Integrating a Multi-Agent Recommendation System into a Mobile Learning Management System*, Proc. of Artificial Intelligence in Mobile Systems 200, October 12, Seattle, USA.

[15] Rosaci D., Sarné G. (2010) *Efficient Personalization of E-learning Activities Using a Multi-Device Decentralized Recommender System*, Comput. Intell., 26/2010, 121-141.

[16] Zaïane O.R. (2002) *Building a Recommender Agent for E-learning Systems, Proc. of the 7th Int. Conf. on Computers in Education*, Auckland, New Zeland, 55-59.

[17] Zakrzewska D. (2011) *Building Context-Aware Group Recommendations in E-learning Systems*, LNAI 6922, Jędrzejowicz P., Nguyen N.T., Hoang K. (eds.): Computational Collective Intelligence - Technologies and Applications. ICCCI 2011, Part I, pp. 132-141.

[18] Dey A.K. (2001) *Understanding and Using Context*, Pers. Ubiquit. Comput., 5/2004, 4-7.

[19] Das M.M., Chithralekha T., SivaSathya S. (2010) *Static Context Model for Context Aware E-learning*, International Journal of Engineering Science and Technology, 2/2010, 2337–2346.

[20] Viola S.R., Graf S., Kinshuk, Leo T. (2007) *Investigating Relationships within the Index of Learning Styles: a Data Driven Approach*, Interactive Technology & Smart Education, 4/2007, 7–18.

[21] Felder R.M., Silverman L.K. (1988) *Learning and Teaching Styles in Engineering Education*, Eng. Educ., 78/1988, 674–681.

[22] Index of Learning Style Questionnaire, http://www.engr.ncsu.edu/learningstyles/ilsweb.html

[23] De Marsico M., Levialdi S. (2004) *Evaluating web sites: exploiting user's expectations*, Intern. Journal of Human-Computer Studies, 60/2004, 381-416.

[24] Zakrzewska D., Wojciechowski A. (2008) *Identifying students usability needs in collaborative learning environments*, Proc. of 2008 Conference on Human System Interaction, Cracov, 862-867.

[25] Zakrzewska D. (2008) *Validation of Cluster Analysis Techniques for Students' Grouping in Intelligent E-learning Systems*, Proceedings of 14th International Congress of Cybernetics and Systems of WOSC, Wroclaw, Poland, 893-901.

[26] Witten I.H., Frank E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Morgan Kaufmann Publishers, San Francisco, CA.