Personalized Data Mining for University Curricula

Rainer Knauf[1], Yoshitaka Sakurai[2], Kohei Takada[2], Setsuo Tsuruta[2]
[1]Faculty of Computer Science and Automation, Ilmenau University of Technology, Ilmenau
[2]School of Information Environment, Tokyo Denki University, Tokyo
rainer.knauf@tu-ilmenau.de

Abstract: *The authors developed a modeling system for university learning processes, which aims at evaluating and refining university curricula to reach an optimum of learning success in terms of a best possible grade point average (GPA). This is performed by applying an Educational Data Mining (EDM) technology to former students curricula and their degree of success (GPA) and thus, uncovering golden didactic knowledge for successful education. We used learner profiles to personalize this technology. After a short introduction to this technology, we discuss the result of a practical application and draw conclusions. The particular contribution of this paper is a "lazy" strategy of mining with data, which is really available without making "guesses" what they mean (profiles). In particular, we utilize the educational history of the students and vocational ambitions for student modeling.*

Keywords: modeling learning processes, storyboarding, educational data mining, learner profiling

1. Introduction

The application of didactic skills in teaching situations is not formally modeled for use in academic education. To make didactic design explicit, we developed a formal modeling approach called storyboarding. Storyboarding is setting the stage to apply Knowledge Engineering Technologies to verify, validate, and refine the didactics behind a learning process. Moreover, didactics can be refined according to revealed weaknesses and proven excellence based on students' storyboard paths and their related learning success in terms of their achieved grade point average (GPA). This technology is adaptive in terms the mining results to both (1) the educational history of the considered student, which improves the students characteristics (“lazy profile”) more and more and (2) the data base whose data is dynamically updated by the more and more students' study results for the DM technology. According to the systematic of Educational Data Mining (EDM) in (Baker & Yacef, 2009), our approach belongs to two categories, namely (1) prediction by classification (see section 2) and (2) clustering (see section 3). The paper is organized as follows. Section 2 is an introduction to storyboarding and introduces the EDM technique for storyboard paths. Section 3 introduces the personalization approach. Finally, we present evaluation results (section 4) and conclude the paper in section 5.

2. Educational Data Mining on Storyboard Paths

A storyboard is defined as a nested graph and may be seen as a model of an anticipated reception process that is interpreted as follows. *Scenes* denote a non-decomposable learning activity that can be implemented in any way. It can be the presentation of a (media) document, opening a tool that supports learning (URL or e-learning system) or an informal activity description. Graphically, it is represented by a rectangle. In case it is a media content, double click on it results in opening the related media file or URL. *Episodes* are defined by their sub-graph. Graphically, it is represented as a rectangle with double vertical lines. A double click on it will cause jumping into the related sub-graph; a click on the End-Node of the sub-graph will result in jumping back to the related episode node in the super-graph. A *Start Node* of a (sub-) graph defines the starting point of a legal graph traversing. An *End Node* of a (sub-) graph defines the final target point of a legal graph traversing. *Edges* denote transitions between nodes and may be (single- or bi-) colored. The outgoing edge must have the same color as the incoming edge by which the node was reached. Thus, the colors express the interdependence between incoming and outgoing edges of a node. For detailed information on this concept, see (Knauf et al., 2010).

An objective of storyboarding is to use knowledge engineering technologies on the (semi-) formal process models. In particular, we aim at inductively “learning” successful storyboard patterns and recommendable paths. It is performed by analyzing paths former students went through the storyboard.

2.1. Data Preprocessing

In a pre-processing step, the episodes are recursively replaced by the paths in their sub-graphs until there is no episode any more in the path as illustrated in Figure.

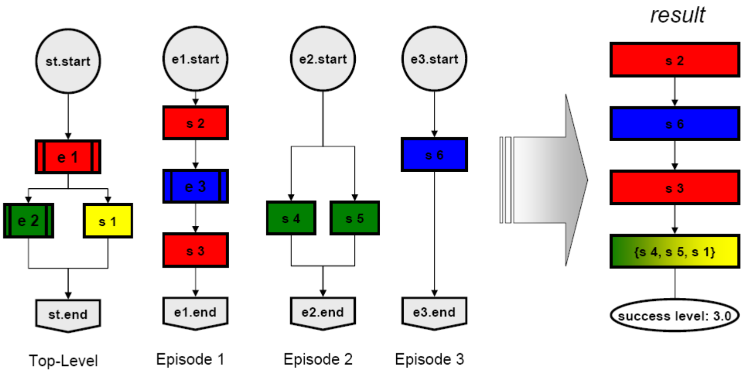
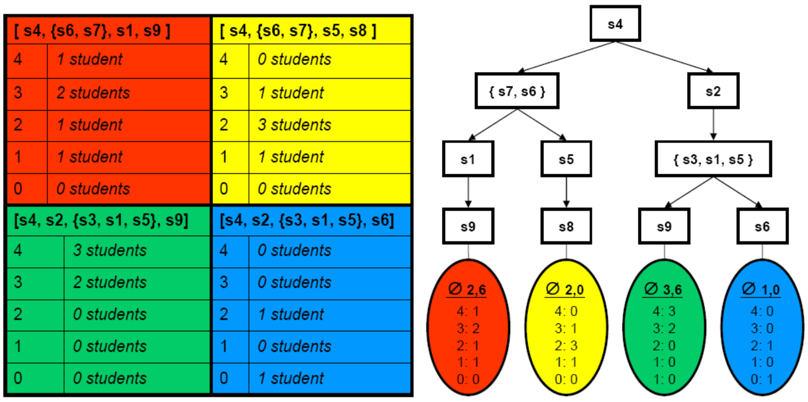


Figure 1. Data Preprocessing Example

2.2. Decision Tree Construction

After this “flattening”, the paths are available as sequences of atomic nodes or sets of atomic nodes and end with a GPA label node. The decision tree is based on the concept of bundling common starting sequences of the various paths to a node of the tree. Different subsequent following (next) nodes of the paths will result in different sub-trees right below the actual root on the last node of the common starting sequence. This continues for each lower level sub-tree accordingly.

The decision tree construction is illustrated in Figure . The final nodes of the paths are followed by a label-node. Label-nodes contain a list of GPAs that students received after going through this path along with the average GPA of all students, who went this way. Each GPA is along with the number of occurrences (the number of students reaching this GPA). The average GPA serves as an estimate of GPA for future students who plan to go through the same path.



[s4, {s6, s7}, s1, s9]		[s4, {s6, s7}, s5, s8]	
4	*1 student*	4	*0 students*
3	*2 students*	3	*1 student*
2	*1 student*	2	*3 students*
1	*1 student*	1	*1 student*
0	*0 students*	0	*0 students*
[s4, s2, {s3, s1, s5}, s9]		**[s4, s2, {s3, s1, s5}, s6]**	
4	*3 students*	4	*0 students*
3	*2 students*	3	*0 students*
2	*0 students*	2	*1 student*
1	*0 students*	1	*0 students*
0	*0 students*	0	*1 student*

Figure 2. Decision Tree Construction

2.3. Decision Tree Utilization

If a submitted path is completely represented in the decision tree, the GPA estimation is easily done through presenting the content of this label. In the other case, namely, if a student submits a curriculum plan that is not represented in the decision tree, the most similar sub-path in the decision tree will be identified. Similarity refers to the number of same course sets in sequence, which the path has in common with a path represented in the tree. This similarity measure s is in the range $0 \leqq s \leqq 1$. In the worst case, there is no node in common with any path in the tree (s = 0) and in the best case, the submitted path is completely represented in the tree (s = 1). This is performed by simultaneously traversing the path's course sequence and the decision tree down from the root until (1) the path is finished or (2) there is a “next node” in the path that is different from all “next sub-tree roots”.

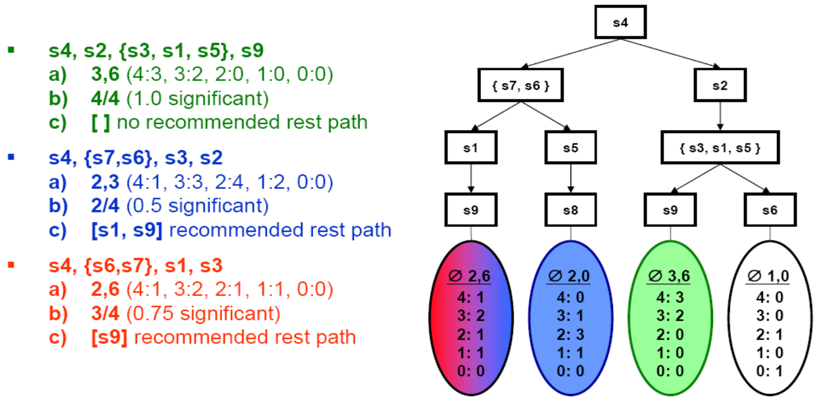
In the first case, the related label node at its leaf position provides the desired GPA estimation. In the latter case, the label node of the current tree position provides the desired information. In such a case, one may be interested in suggestions to modify the submitted path in a way that the chances for a higher GPA reach an optimum or become the highest in their value. Thus, it is suggested to exchange the submitted remaining path for the most successful alternative remaining path with the best GPA among the ones in the decision tree. For this purpose, we supplement the estimated GPA with the most successful remaining path starting at the last node of the tree traversing. We provide this optimal supplement along with its achievable GPA, if this optimum is better than the GPA estimation of the submitted path.

Also, the user is informed of the degree of similarity of his submitted path and the one found in the decision tree. We call this similarity significance and compute it as the number of nodes in sequence that are common in the submitted path and the decision tree, related to the entire length of the path.

More concretely, the path's GPA estimation by using is performed as follows:

1. If the submitted path is completely represented in the tree, (a) the GPA estimation will be given by providing a GPA label, (b) the significance of this estimation is 1, because it is based on information of the complete submitted path, and (c) the recommended rest path is empty, because the decision tree does not contain a supplemental path, which leads to tan GPA improvement.
2. If the submitted path is partly represented in the tree, (a) the GPA is computed by merging the GPA labels of all sub-trees starting from the last node that have both the submitted path and a related path in the tree in common, (b) the significance is computed as the number of common nodes divided by the total number of nodes in the path, and (c) the recommended rest path is the best rated path in the tree after the end node of submitted path.

The utilization is illustrated in Figure . Here are three example cases of utilization, namely (1) the path under evaluation is completely in the decision tree (green path), (2) the path is partly in the decision tree and several alternative supplements are available (blue path), and (3) the path is partly in the decision tree and only one way to continue it is represented in the decision tree (red path). In the first case (green) (a) the information on the average success of former students, who went this path is provided along with the distribution of marks, (b) a significance of 1 is provided, which means there was information on the complete path available, and (c) no recommendation of the best way to continue this path can be given. In the second case (blue) (a) the information on the average success of former students, who also went this path is provided along with the distribution of marks, which is calculated by merging both leaf nodes of the alternative sub-trees after the node *{s7, s6}*, (b) a significance of 0.5 is provided, which means there was information on half of the path available, and (c) a recommendation of the best way to continue this path is given. In the third case (red) (a) the information on the average success of former students, who also went this path is provided along with the distribution of marks, (b) a significance of 0.75 is provided, which means there was information on 75 % of the path available, and (c) a recommendation of the "best" way to continue this path is given by providing the only one rest path presented in the tree.



- **s4, s2, {s3, s1, s5}, s9**
  - **a)** **3,6** (4:3, 3:2, 2:0, 1:0, 0:0)
  - **b)** **4/4** (1.0 significant)
  - **c)** **[ ]** no recommended rest path
- **s4, {s7,s6}, s3, s2**
  - **a)** **2,3** (4:1, 3:3, 2:4, 1:2, 0:0)
  - **b)** **2/4** (0.5 significant)
  - **c)** **[s1, s9]** recommended rest path
- **s4, {s6,s7}, s1, s3**
  - **a)** **2,6** (4:1, 3:2, 2:1, 1:1, 0:0)
  - **b)** **3/4** (0.75 significant)
  - **c)** **[s9]** recommended rest path

Figure 3. Decision Tree Utilization

3. Personalization by Student Modeling

Our mining results would be more significant, if individual properties, talents, and preferences are considered as well. For example, some students are more talented for analytical challenges, some are more successful in creative or composing tasks, and others may have an extraordinary talent to memorize a lot of factual knowledge. Consequently, we need to include individual student profiles to avoid lavishing upon the students, suggestions that don't match their individual preferences and talents, which are derived from their educational history and additional data that describes individual characteristics. In (Knauf et al., 2009) we introduced an approach of personalized EDM by deriving estimating intellectual traits (Gardner, 1993) and learning styles (Felder & Silverman, 1988). Unfortunately, we could not obtain sufficient data for this kind of learner profiles. Therefore, we refrained from deriving an explicit model. Instead, we shifted strategy from an "eager" strategy of holding an explicit model towards a "lazy" strategy of mining with data, which is really available, holds empirically, and is not a result of "guesses" about the students' general characteristics. In particular, we utilize the educational history of the students and vocational ambitions for student modeling. This new issue is the focus of this paper.

3.1. Modeling Educational History

As educational data of a student, we have for each completed semester (1) the subjects taken, (2) their number of units, (3) the rating in it reached by the student (S, A, B, C, D, or E), and (4) the related Grade Points (4 for S and A, 3 for B, 2 for C, and 0 for D and E). From this data, we form a set of pairs $p = \{[s_1, l_1], [s_2, l_2], \ldots, [s_n, l_n]\ \}$ with $s = \{s_1, \ldots, s_n\}$ being the set of codes for subjects taken by the student, and l_i being number of grade points achieved by the student in subject with the code s_i. To include pre-university education and other available data, we include four subjects that are part of pre-university examinations, namely (1) *Mathematics, Science (mostly Physics)*, (2) *English and Language Arts (Japanese)*, and (3) three "subjects" that might characterize a student, too: *High School Recommendation*, *Self Recommendation*, and some (proven and checked by the university administration) *Special Skills*.

3.2. Modeling Vocational Ambitions

For the self-estimation of vocational ambitions, the student is asked about his/her preferred field in his/her future job. The student is asked, "What kind of job he/she prefers", namely (1) topical work, (2) administrative work, (3) research work, and/or (4) teaching work, and can select one of these categories. Then, the student selects topical fields, in which he/she looks for such a position as characterized above. This is performed in 2 two level manner. He/she is requested to rate on a scale ranging from 0 to 4. According to the faculty structure at the School of Information Environment (SIE) of the Tokyo Denki University, they are firstly asked to rate three major fields. For those of these major fields, which received a rate different from zero, he/she is asked to rate more precisely among three subfields. As a result, we have up to 13 pairs [ambition, level], namely up to 4 characteristics and up to 9 topical fields.

3.3. Derived Student Model

Initially, our student profile is a vector of pairs *[profile item, level]*, namely up to 5 pairs *[subject, success level]*, which represent the educational history and up to 13 pairs *[vocational ambition, level]*, which represent vocational ambitions. All these pairs have the same range of their levels, namely 0 through 4. As soon as new learning success results become available, the profile is extended by the results in the courses the student has taken so far and gains more and more items after each semester according to the number of courses the student has taken. For a personalized GPA prediction, we compute a degree of similarity between the educational history of a student under estimation and that of the students in our data base. Similarity of profiles means similar relations in-between its components, which is expressed by the Cosine Coefficient. To compute a similarity measure $sim(p_1,p_2)$ for two profiles $p^1 = \{[s_1^1, l_1^1], [s_2^1, l_2^1], ..., [s_n^1, l_n^1]\}$ and $p^2 = \{[s_1^2, l_1^2], [s_2^2, l_2^2], ..., [s_m^2, l_m^2]\}$, we form two k - dimensional ($k \leqq n + m$) vectors $\overrightarrow{p^1} = [l_1^1, l_2^1, ..., l_k^1]$ and $\overrightarrow{p^2} = [l_1^2, l_2^2, ..., l_k^2]$ by

- merging the subject sets $s^1 = \{s_1^1, s_2^1, ..., s_n^1\}$ and $s^2 = \{s_1^2, s_2^2, ..., s_n^2\}$ \$ s^2 = \\{s_1^2, s_2^2, towards $s = s^1 \cup s^2 = \{s_1, s_2, ..., s_k\}$ with $k \leqq n + m$,
- adopting the l_i^j from the original profiles, if $s_i^j \in \Pi_1(p^j)$,
- setting $l_i^j := 0$, if the considered subject is not in the related profile ($s_i^j \notin \Pi_1(p^j)$), and
- using the Cosine coefficient as similarity measure $sim(p^1, p^2) = \cos(\overrightarrow{p^1}, \overrightarrow{p^2})$.

In our new approach that includes the student profiles, construct the decision tree exclusively from students with profiles that are most similar to the one under evaluation. To compose the subset $s^{sim} \subseteq s$ of most similar students to a student under evaluation with a student profile p^{eval}, we state a portion (a percentage *prc*) of students, whose paths are most similar to the submitted ones. This way,

(1) the estimation of success chances in terms of an likely to achieve GPA is based on individual preferences, talents, and weaknesses, and

(2) the suggestion of a remaining curriculum path (subjects recommended to optimize the success of study) is adapted to individual properties, because it is calculated on the base of examples with a similar profile

by applying the technology to groups of students with similar attributes, behaviors, outcomes only.

In contrast to our former profiling approach (Knauf et al., 2009), this profiling approach is quite dynamic and improving over time. Subjects currently taken will be history in the next semester and their results are useful information for a more precise profiling.

4. Evaluation of the Approach

We collected 186 individual storyboard paths of students, who studied Information Environment at the School of Information Environment of the Tokyo Denki University from 2005 till 2009. After studying all the samples and organizational material rules to compose a curriculum, we chose a compact data representation by coding the particular subjects and the particular students. By using subject codes 1-155 and student IDs 1-186, we composed a complete decision tree from the 186 samples.

To make sure that identical starting sequences of semester curricula really end up in the same path, the decision tree is well sorted: (1) the subject sequence within a semester is sorted by ascending subject codes and (2) the students samples are sorted by the code lists, which are, compared element by element, ascending, too. We adopted this technology from a similar technology used in Data Mining to efficiently generate association rules.

To evaluate this approach, we used these 186 individual storyboard paths and validated the approach by cross validation with a subset size of 1, i.e. the so called *leave one out* approach. In a number of cycles equal to the number of samples, we removed one path from the complete decision tree and used this sample to check the remaining decision tree. As a result, we received a list of all the samples along with their original (real) GPA and the GPA as estimated by the EDM technology. The mean of the difference between both was 0.43 with a standard deviation of 0.30. The linear correlation coefficient between the estimated and the real GPA was 0.58.

Of course we wanted to know which circumstances promise a good estimation and which do not. Unfortunately, due to data privacy protection, we did not have other data of the students such as age, sex, or family status, although all the data was anonymous (each student was just a number in our study). The only thing we could analyze is, whether or not there is a correlation between the proficiency level (GPA) and the quality of the results of our technology. We expected the technology to be worse in low level GPA and becoming better and better with increasing GPA, but got a surprising result.

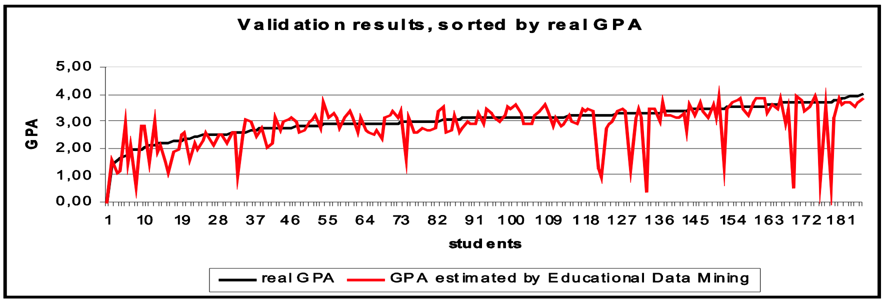


Figure 4. Validation Histogram, sorted by PPA

As shown in Figure , our expectation about low level GPA students was apparently not wrong and we mistook in both directions, i.e. over- and underestimated their GPA. For the majority of students, who have a GPA around three, we achieved mostly good results. But for very high proficiency students with a GPA of 3.5 or higher (these were 37 out of the 186 students) the risk that our estimation is far from the truth, was growing again. For all students except the 20 best, our results are much closer to the truth (mean of the difference 0.36 with a standard deviation of 0.44, correlation 0.67) than for the 20 best students (mean of the difference 0.65 with a standard deviation of 1.04, correlation 0.02).

Currently, we are discussing, what additional data we could include into our EDM technology to improve the result especially for this kind of students.

5. Conclusions

The paper introduced an intelligent system for modeling and supporting academic educational processes. The intelligence of the system is performed by EDM over a semi-formal model of the process by means of storyboarding. In this way, optimal didactic success patterns with proven excellence can be inferred, which may have never been revealed by conventional (non-intelligent) methods of analyzing students' educational data. It consists of mining a decision tree and applying this decision tree to curricula planned by current students. Thus, students receive an estimation of the GPA along with a recommendation to supplement it to reach optimal success. This technology is adaptive in terms of computing the mining results due to both (1) the educational history of the considered student and (2) the data base whose data is dynamically updated by the students' study results for the DM technology.

Here, an approach to include individual learner profiles was introduced. The profiling concept initially uses the per-university educational history and is dynamically extended by the students' university study results. In this way, the students are grouped into groups with similar attributes, behaviors, and outcomes. The original general (non-individual) method, namely, the proposed DM method, is applied to students with profiles of a high degree of similarity to the student under consideration.

Due to a lack of data, we were not able to implement our initial learner profiling approach, which aimed at an explicit acquisition of intelligence traits and learning styles. Instead, we shifted strategy from an "eager" strategy of holding an explicit model towards a "lazy" strategy of mining with data, which is really available, holds empirically, and is not a result of "guesses" about the students' general characteristics. In particular, we utilize the educational history of the students and vocational ambitions for student modeling. This new issue is the focus of this paper. Our upcoming work is focused in a new evaluation experiment, which includes the new profiling concept.

A feasibility study showed its usefulness of the system. It supported students' dynamic learning activities by features that help to overview, verify, refine, and optimize their own individual curricula. The effect has been validated by cross-validation with about 200 students' records.

As future work, we will investigate the reasons for the weaknesses of the approach for with the very good and very bad students. Further, we think about introducing a correlation analysis between courses (including negative correlations) to suggest appropriate courses based on a student's result so far.

References

Baker, R. S. J. D., & Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 2009.

Felder, R. M., & Silverman, L. K. (1988) Learning and teaching styles in engineering education. *Engineering Education*, 78(7), pp.\ 674-681.

Gardner, H. (1993) *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.

Knauf, R, Sakurai, Y., Takada, K., & Dohi, S. (2009) Personalized Curriculum Composition by Learner Profile Driven Data Mining. *Proc. of the 2009 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2009)*, ISBN 978-1-4244- 2794-9, San Antonio, TX, USA, pp.\ 2137-2142.

Knauf, R., Sakurai, Y., Tsuruta, S., & Jantke, K. P. (2010) Modeling Didactic Knowledge by Storyboarding, *Journal of Educational Computing Research*. (42)4, ISSN: 0735-6331 (Paper) 1541-4140 (Online), Baywood Publishing Company Inc, pp. 355-383.

Sakurai, Y., Dohi, S., Tsuruta, S., & Knauf, R. (2009) Modeling Academic Education Processes by Dynamic Storyboarding. *Journal of Educational Technology & Society*, vol. 12, ISSN 1436-4522 (online) and 1176-3647 (print), pp. 307-333.