VALIDATION OF A DATA MINING METHOD FOR OPTIMAL UNIVERSITY CURRICULA

R. Knauf[*]

Ilmenau University of Technology
Faculty of Computer Science
and Automation
PO Box 100565, 98684 Ilmenau
Germany

Y. Sakurai, K. Takada, S. Tsuruta

Tokyo Denki University
School of Information Environment
2-1200 MuZai Gakuendai
Inzai, Chiba, 270-1383
Japan

ABSTRACT

The paper deals with modeling, processing, evaluating and refining processes with humans involved like learning. A formerly developed concept called storyboarding has been applied at Tokyo Denki University to model the various ways to study at this university. Along with this storyboard, we developed a data mining technology to estimate success chances of curricula. Here, we introduce a validation method for this technology and its results. Further, we discuss chances to improve these results by implementing a formerly introduced learner profiling concept that represents the students' individual properties, talents and preferences for personalized data mining.

Index Terms— modeling learning processes, storyboarding, data mining, validation

1. INTRODUCTION

Learning systems suffer from a lack of an explicit and adaptive didactic design. University education is especially effected by this lack, because university professors are not necessarily educational experts. One way of didactic support is providing a modeling concept for didactic design, which allows the anticipation of the learning processes.

An explicit formal didactic design provides a firm basis to verify and validate the didactics behind a learning process by knowledge engineering techniques such as machine learning and data mining. A modeling concept called storyboarding [1] has been developed formerly as a means of modeling learning processes. Besides providing didactic support, this semi-formal model is setting the stage to apply knowledge engineering technologies to verify and validate the didactics behind a learning process. The verification may include both logical consistency issues and formally to check didactic issues. According to different learning and teaching preferences, it includes alternative paths and possible detours if certain concepts to be learned need reinforcement. Using modern media technology, a storyboard also plays the role of a server that provides the appropriate content material.

By storyboarding, didactics can be refined according to revealed weaknesses and proven excellence. Successful didactic patterns can be explored by applying data mining techniques to the various ways students went through a storyboard and their related success. As a result, future instructors and students may utilize these results by preferring those ways through a storyboard, which turned out to be the most promising ones. In [2], a data mining technology, which allows students to utilize mined "experience" of former students to compose curricula with an optimal success chance, is introduced.

However, so far we did not have a practically proven significance, that this method is appropriate. The basic problem so far was the collection of data, which has to be accumulated during a complete undergraduate study, which needs a period of four years. Meanwhile, we could gain a significant amount of data to validate the technology.

The paper is organized as follows. Section 2 introduces the storyboard concept including the present state of the current development. Section 3 provides an overview on our data mining technique to compose optimal curricula for university studies. In section 4, we describe the available data. Section 5 introduces the validation technology and provides its results. In section 6, we outline a refinement of the technology and section 7 summarizes the paper.

2. STORYBOARDING

Our storyboard concept was introduced in [1] und later refined (see [2] for the latest version). A storyboard is a nested hierarchy of directed graphs with anno-

[*] This author performed the work while at Tokyo Denki University and was sponsored by the Japan Society for the Promotion of Science (JSPS) with an Award-Fellowship for Rainer Knauf (Fellow's ID S-08742) and the Research Institute for Science and Technology of Tokyo Denki University.

tated nodes and annotated edges. Nodes are *scenes* or *episodes*. *Scenes* are not further structured, *episodes* have a sub-graph as its implementation. Also, there is exactly one *start node* and one *end node* in each graph. Edges specify transitions between nodes and may be single-color or bi-color. Nodes and edges can carry attributes.

A storyboard 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, e.g. by the presentation of a (media) document, opening a tool that supports learning (an URL or an e-learning system) or an informal activity description. *Episodes* are defined by their sub-graph. *Graphs* are interpreted by the paths, on which they can be traversed.

A *start node* of a graph defines the starting point of a legal graph traversing. An *end node* of a graph defines the final target point of a legal graph traversing.

Edges denote transitions between nodes. There are rules to leave a node by an outgoing edge, namely (1) The outgoing edge must have the same color as the incoming edge by which the node was reached and (2) If there is a condition specified as the edge's key attribute, this condition has to be met for leaving the node by this edge. So the colors express the dependence of ways leaving a node from the way of arriving there.

Key attributes of nodes specify application driven information, which is necessary for all nodes of the same type, e.g. actors and locations. *Key attributes of edges* specify conditions, which have to be true for traversing on this edge. Free attributes specify whatever the storyboard author wants the user to know: didactic intentions, useful methods, necessary equipment, e.g. For further information, the reader may see [3] or [4].

3. CURRICULUM VALIDATION BY DATA MINING

A basic objective of storyboarding is to use knowledge engineering technologies on the (semi-) formal process models [3] [4].

In particular, we aim at inductively "learning" successful storyboard patterns and recommendable paths. This is some sort of meta-learning, i.e. the learning of learning knowledge. It is performed by an analysis of the paths where former students went through the storyboard [2].

To show the feasibility and benefit of high level storyboarding for its qualified assistance of students suffering from the "jungle of opportunities and constraints" in university education, we developed a simple prototype storyboard for curricula of a university study.

This prototype is used to validate curricula, which are created or modified by the students in advance of their study [4][2] based on the success of former students, who went a similar path through their study.

For this purpose, we introduced a concept to estimate success chances of curricula, which are composed by students at the School of Information Environment of the Tokyo Denki University in their curriculum planning class in the first semester. Along with the estimation, the students also receive (1) a significance of the provided estimation statement (according to the sufficiency of the available data) and (2) a recommendation for modifications of their plan with respect to an optimal success chance.

For such curricula we developed a data mining technique, which is applied to storyboard paths that (former) students went. Based on these examples, the success chance of intended paths can be estimated [2].

The data mining technique is applied to the paths of students through a storyboard, which anticipates possible ways through a complete study.

In a pre-processing step to determine the paths, the individually visited items (episodes and scenes) in the storyboard graph-hierarchy are "flatten down" to a big graph that contains scenes only. This is performed by systematically replacing episodes by the individually visited items of the episode's related sub-graph.

In the granularity of this storyboard application, a scene is a course that holds over one semester. As a result, we have a linear list of course sets, in which each list item is the set of courses that the student took in the subsequent semesters.

The technique consists of two steps, namely (1) constructing a decision from the examples of former students and (2) applying this decision tree to the planned curricula.

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. If there are different paths with a common starting sequence from the root to the actual root different in the next (subsequent) nodes, related sub-trees will be established.

The utilization or application of this decision tree is performed as follows.

If a submitted path is already represented in the decision tree, the prediction or estimation is very easily done through presenting the average Grade Point Average (average of a numeric performance metric of a student over all subjects, weighted by the number of each subject) that students gained, who went exactly this paths, too.

In the other case, the longest leading (starting and its succeeding) part in common with the path representing the submitted curriculum plan will be identified and

code	subject
1	Advanced Project A
2	Advanced Project B
3	Agent Technology
⋮	⋮
155	Workshop

Table 1. Subject list

the average GPA of all students' paths in the sub-trees that start from that point, will be presented as a success estimation. Additionally, the degree of similarity and a recommended change of the submitted path will be presented. T he data mining technology is described more detailed in [2].

4. DATA PREPROCESSING

We collected 188 individual storyboard paths of students, who studied Information Environment at the School of Information Environment of Tokyo Denki University from 2005 till 2009.

From these samples, we removed two samples of students, who joint the university after taking several semesters elsewhere, because their marks were derived by recognition of marks received in similar subjects at another university. This led to 186 samples.

After collecting and studying all the samples and organizational material rules to compose a curriculum, which was available in Japanese only, we chose a compact data representation by coding the particular subjects and the particular students. Table 1 shows an extract from the subject coding list.

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, which is usually performed in data mining for item lists to efficiently generate association rules.

Figure 1 shows an extract of the decision tree composed by all the samples. For each student (coded by his/her ID),

- each semester (columns s, with yellow-brown background),
- the subjects (courses, columns c with light green background),
- their number of units (columns u with light yellow background) and
- the achieved results (with light blue background), i.e. the mark (columns m: S, A, B, C, D, or E) and the number of grade points (columns GP: 4, 3, 2, or 0)

are listed up.

The last row contains a weighted (by the number of units) grade point average GPA, which quantifies the degree of success in the study. Again, both the subject lists of the students within a semester and the complete students' samples (which are lists of lists), are sorted by subject code. The bars between the paths show, up to which semester the curricula of adjacent students are identical (circles) respectively from which semester they are different from each other (bullets). Thus, the grey bars separate the sub-trees from each other.

The entire table has 42 columns and 1616 rows. Figuratively spoken, the table illustrates the decision tree in a horizontal direction wit the root being on the very left hand side and the leaves being on the very right hand side. The grey bars separate sub-trees from each other.

Before applying the validation technology, we found some "exotic samples" of students, who are not representative. This applies to those students, who never finished their study (as this was the case with students 8, 11, 59, 97, 113, 118, 121 and 153) and removed them because of incomplete data, i.e. 177 samples left. As a "learning curve", in future validations, we will leave at least those "dead end" paths in the set, which are caused by a lack of performance.

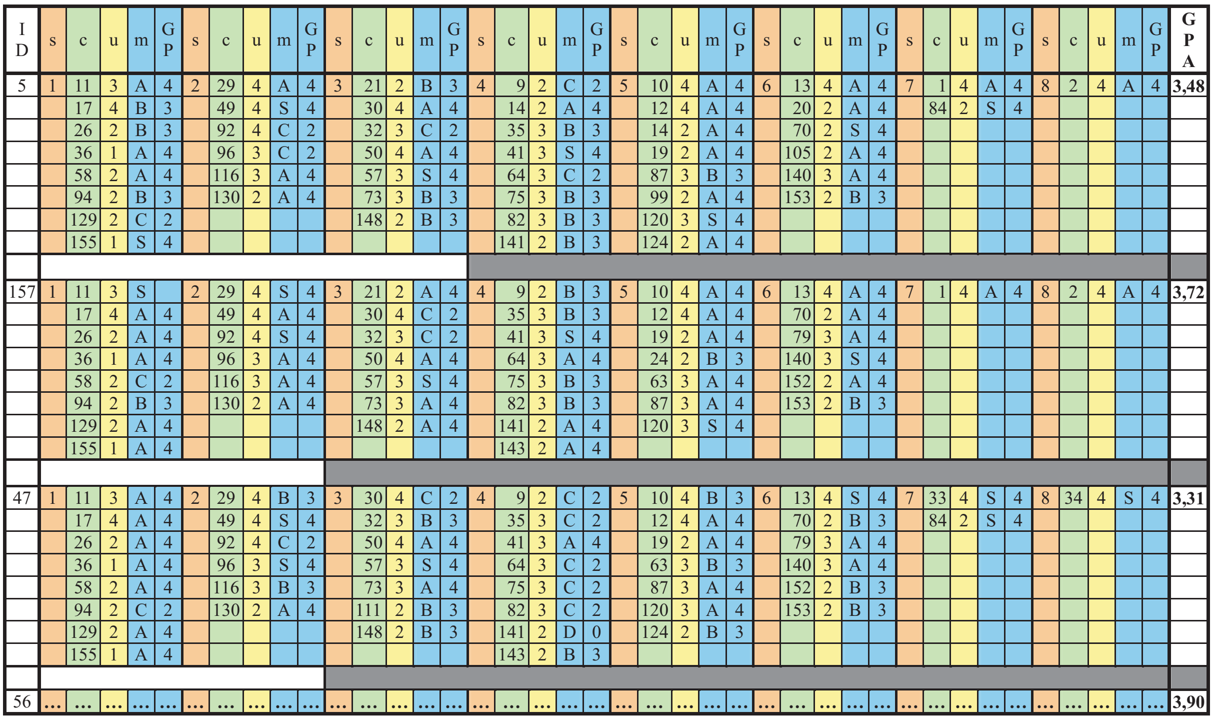
Our validation technology uses an example set to construct a decision tree and a test set to check its performance. Both the example set and the test set are recruited from the given samples.

Those storyboard paths, which are unique and do not have anything in common with any other path, are not appropriate for such a technology, because the test set origins from the same source of data. If the test set contained samples that do not have anything in common with any path of the decision tree, any data mining can not really work because of missing data.

In practice, our data mining technology degenerates to merge all paths of the decision tree and provides the average degree of success of all former students.

Since this is not really a result of data mining, we excluded such paths, which led us to 104 remaining paths, which are used to validate the technology.

For practical use in the success estimation of new paths submitted by students, however, we kept these 73 "lonely" paths, of course, because new paths may be similar to them as well. In fact, any new path is ”lonely” when somebody goes it the first time, before it may gain popularity and grow evolutionary towards a sub-tree.



ID	s	c	u	m	GP	s	c	u	m	GP	s	c	u	m	GP	s	c	u	m	GP	s	c	u	m	GP	s	c	u	m	GP	s	c	u	m	GP	s	c	u	m	GP	GPA
5	1	11	3	A	4	2	29	4	A	4	3	21	2	B	3	4	9	2	C	2	5	10	4	A	4	6	13	4	A	4	7	1	4	A	4	8	2	4	A	4	**3,48**
		17	4	B	3		49	4	S	4		30	4	A	4		14	2	A	4		12	4	A	4		20	2	A	4		84	2	S	4						
		26	2	B	3		92	4	C	2		32	3	C	2		35	3	B	3		14	2	A	4		70	2	S	4											
		36	1	A	4		96	3	C	2		50	4	A	4		41	3	S	4		19	2	A	4		105	2	A	4											
		58	2	A	4		116	3	A	4		57	3	S	4		64	3	C	2		87	3	B	3		140	3	A	4											
		94	2	B	3		130	2	A	4		73	3	B	3		75	3	B	3		99	2	A	4		153	2	B	3											
		129	2	C	2							148	2	B	3		82	3	B	3		120	3	S	4																
		155	1	S	4												141	2	B	3		124	2	A	4																
157	1	11	3	S		2	29	4	S	4	3	21	2	A	4	4	9	2	B	3	5	10	4	A	4	6	13	4	A	4	7	1	4	A	4	8	2	4	A	4	**3,72**
		17	4	A	4		49	4	A	4		30	4	C	2		35	3	B	3		12	4	A	4		70	2	A	4											
		26	2	A	4		92	4	S	4		32	3	C	2		41	3	S	4		19	2	A	4		79	3	A	4											
		36	1	A	4		96	3	A	4		50	4	A	4		64	3	A	4		24	2	B	3		140	3	S	4											
		58	2	C	2		116	3	A	4		57	3	S	4		75	3	B	3		63	3	A	4		152	2	A	4											
		94	2	B	3		130	2	A	4		73	3	A	4		82	3	B	3		87	3	A	4		153	2	B	3											
		129	2	A	4							148	2	A	4		141	2	A	4		120	3	S	4																
		155	1	A	4												143	2	A	4																					
47	1	11	3	A	4	2	29	4	B	3	3	30	4	C	2	4	9	2	C	2	5	10	4	B	3	6	13	4	S	4	7	33	4	S	4	8	34	4	S	4	**3,31**
		17	4	A	4		49	4	S	4		32	3	B	3		35	3	C	2		12	4	A	4		70	2	B	3		84	2	S	4						
		26	2	A	4		92	4	C	2		50	4	A	4		41	3	A	4		19	2	A	4		79	3	A	4											
		36	1	A	4		96	3	S	4		57	3	S	4		64	3	C	2		63	3	B	3		140	3	A	4											
		58	2	A	4		116	3	B	3		73	3	A	4		75	3	C	2		87	3	A	4		152	2	B	3											
		94	2	C	2		130	2	A	4		111	2	B	3		82	3	C	2		120	3	A	4		153	2	B	3											
		129	2	A	4							148	2	B	3		141	2	D	0		124	2	B	3																
		155	1	A	4												143	2	B	3																					
56	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	**3,90**

Fig. 1. Extract from the decision tree data

5. VALIDATION TECHNOLOGY AND RESULTS

There are several approaches to validate data mining technologies.

The *holdout method* splits the data into a training set and a test set, typically in the ratio 2/3 by 1/3. The data mining technology is applied to the training set and validated with the test set. This method suffers from the fact that it does not use the available data exhaustively. A sample, which is in the test set, is not available for building the model (the decision tree, in our case) and thus, decreases the performance of the model. Thus, some performance features of the data mining technology may not be revealed by such a testing method. The splitting ratio is a trade off between the quality of the model and a trustable statement about the performance of the data mining technology.

Random sub-sampling is a refinement of this method, which is a repeated holdout with various splits of the available data and thus, uses the data a little more exhaustively. However, there is no control on the issue, how often a data object is used for building the model and how often it is used for test.

A more exhaustive utilization of the available data is done by *cross validation*. Here, each data object is used for training with the same frequency and for test exactly once. The data set is split into k equally sized subsets. In k cycles, each subset is used for test, whereas the the other $k-1$ sets is used for training.

stud. ID	GPA	GPA estimation	difference
89	3.40	3.23	0.17
148	3,04	3,26	0,22
179	3,30	3,24	0,06
92	3,55	3,63	0,08
178	3,91	3,40	0,51
164	3,29	3,71	0,42
177	3,52	3,60	0,08
⋮	⋮	⋮	⋮

Table 2. Validation results

The *leave one out* approach is a special case of cross validation with k being the number of data objects and makes the most exhaustive use of the data.

Finally, we used this approach to validate our data mining technology. In 104 cycles, 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 104 samples along with their original GPA and the GPA as estimated by the data mining technology as shown in Table 2. The mean of the difference between both was 0.43 with a standard deviation of 0.30.

Having in mind that this result is just based on a statistical analysis of former students' curricula and their related success, an average error of 0.43 grade points is not too bad and promises remarkable results, when the learner' individual characteristics are also included in the data mining technology.

6. PERSONALIZED DATA MINING AND ITS REALIZATION

Individual learning plans should not only be based on the success of former students who went similar ways. Additionally, individual properties, talents and preferences should be considered.

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 learner profiles to avoid lavishing the students with suggestions that don't match their individual preferences and talents.

In [5], we introduced an approach of personalized data mining. This approach adopts the GARDNER'S theory of multiple intelligences [6] and the learning style model of FELDER and SILVERMAN [7]. The assumption behind this approach is that there is a link between

- typical "competence traits" (according to GARDNER) and subjects that typically challenge the one or other "kind of intelligence" more than others and
- typical teaching methods (according to FELDER and SILVERMAN) and subjects that are typically taught with these methods.

According to [5], the next steps of collecting and processing data to integrate this technology, are (1) the appraisal of the learner profile introduced in [5] for the very best students in each subject, (2) the derivation a typical "success profile" for each subject, (3) the estimation of learner profiles for all students as a (by success degree) weighted average success profile of the subjects they took, and (4) the application of the same technology to the data of "personalized" decision trees for each learner, which are composed by samples of learners, which have a similar learner profile.

The appraisal of the GARDNER - like items in the learner profile can be performed by a questionnaire, which derives an estimation of a human's intelligence distribution by his/her answers on 70 questions. This questionnaire is available to the public in the Internet as a downloadable Microsoft Excel file.[1]

The FELDER-SILVERMAN - like items of the learner profile can be estimated by a questionnaire as well. This questionnaire is also available to the public in the Internet.[2]

attribute	attribute description	value range
d_1	Linguistic intelligence	$0 \le v_1 \le 1$
d_2	Logical-mathematical intelligence	$0 \le v_2 \le 1$
d_3	Musical intelligence	$0 \le v_3 \le 1$
d_4	Bodily-kinesthetic intelligence	$0 \le v_4 \le 1$
d_5	Spatial intelligence	$0 \le v_5 \le 1$
d_6	Interpersonal intelligence	$0 \le v_6 \le 1$
d_7	Intrapersonal intelligence	$0 \le v_7 \le 1$
d_8	Active vs. Reflective style	$0 \le v_8 \le 1$
d_9	Sensing vs. Intuitive style	$0 \le v_9 \le 1$
d_{10}	Visual vs. Verbal style	$0 \le v_{10} \le 1$
d_{11}	Sequential vs. Global style	$0 \le v_{11} \le 1$

Table 3. Derived Learner Profile

We consider both in our model, which is defined as an array of 11 attribute-value pairs that contains 7 intelligence attributes and 4 learning style attributes. Both can be appraised by questionnaires that are available to the public in the web.

To make the dimensions of both sources comparable to each other and see the quantitative relations, we normalized them in a way that they all have the same range of values. The intelligence dimensions rage from 10 to 40. The learning style dimensions range from -11 to +11 (opposite algebraic sign for opposite styles). The normalization can be done by

- $v = result/40$ for the intelligence dimensions according to GARDNER and
- $v = (result + 11)/22$ for the learning style dimensions accodrding to FELDER and SILVERMAN.

Finally, our learner model looks as shown in Table 3.

However, it turned out to be very hard to find former students, who are still accessible and, moreover, willing to fill in such questionnaires to obtain their learner profiles. Our students are very sensible in respecting privacy and, vice versa, in expecting the same respect from others. Since answers to the questions in the questionnaire may reveal some private issues, it is hard to ask them to answer these questions.

However, there are some students, who we dare to ask for filling in the questionnaires because they had a quite confidential relation to the one or other professor, but these students are not necessarily the best ones.

Therefore, steps one and two of this plan need to be changed. To infer a typical "success profile" of a subject, we can collect the questionnaire answers be some student, which are not necessarily the best ones.

Thus, we modified the approach of computing an "average profile" of the best students towards a

[1] see http://www.businessballs.com/howardgardnermultiple...
...intelligences.htm

[2] see http://www.engr.ncsu.edu/learningstyles/ilsweb.html

”weighted average profile” of all available students, who took part in a particular subject.

Let $L(s)$ be the set of learners, who took part in the subject s and for who a learner profile can be composed from the questionnaires’ answers. So for each learner $l^i \in L(s)$, $i = 1...|L(s)|$, a learner profile $p(l^i) = [d_1^i, d_2^i, \cdots, d_{11}^i$ is available. Let

$$succ_s^i = \begin{cases} 1.00 & \text{, if } l^i \text{ received in subject } s \text{ mark } S \\ 0.80 & \text{, if } l^i \text{ received in subject } s \text{ mark } A \\ 0.60 & \text{, if } l^i \text{ received in subject } s \text{ mark } B \\ 0.40 & \text{, if } l^i \text{ received in subject } s \text{ mark } C \\ 0.20 & \text{, if } l^i \text{ received in subject } s \text{ mark } D \\ 0.00 & \text{, if } l^i \text{ received in subject } s \text{ mark } E \end{cases}$$

be the success degree of the learner $l1i$ in subject s.

By using this success degree as a weight factor, the “typical success profile” of a subject s can be computed as

$$p(s) = \frac{1}{\sum\limits_{i=1}^{|L(S)|} succ_s^i} \begin{pmatrix} \sum_{i=1}^{|L(s)|}(succ_s^i * d_1^i) \\ \sum_{i=1}^{|L(s)|}(succ_s^i * d_2^i) \\ \vdots \\ \sum_{i=1}^{|L(s)|}(succ_s^i * d_{11}^i) \end{pmatrix}$$

This calculation has to be done for each subject separately and the set of “most successful students” differs from subject to subject, of course. The idea behind is to mine a “typical success profile” for each subject separately.

After performing these computations, steps three and four can be conducted as planned originally and described in [5]. As a result of processing this additional data in the way sketched above, we expect a remarkable improvement the performance compared to the results presented in section 5.

7. SUMMARY AND OUTLOOK

The research reported here is focused on modeling, processing, evaluating and refining processes with humans involved like learning. A formerly developed concept called storyboarding is briefly introduced.

Along with a storyboard application, we developed a data mining technology to estimate success chances of curricula, which are composed by students. So far, there was no practical significance for the performance of this technology.

The basic problem so far was the collection of data, which has to be accumulated during a complete undergraduate study of, which needs a period of four years. Meanwhile, we could gain a significant amount of data to validate the technology.

By cross validation with the available data, we could empirically show performance of our data mining technology.

However, the currently implemented way of statistically analyzing all former students’ curricula ignores the fact that the success chance heavily depends on individual properties.

A formerly developed approach to validate curricula personalized by building the decision tree based on former students with a similar learner profile only, was refined here. This was necessary, because the required personal data is not available.

As a result of practically implementing this refined approach, we expect a remarkable improvement of these results.

8. REFERENCES

[1] K.P. Jantke and R. Knauf, “Didactic design though storyboarding: Standard concepts for standard tools,” in *Proc. of 4th Int. Symposium on Information and Communication Technologies, Workshop on Dissemination of e-Learning Technologies and Applications, Cape Town, South Africa.* 2005, ISBN 0-9544145-6-X, pp. 20–25, New York: ACM Press.

[2] R. Knauf, R. Böck, Y. Sakurai, S. Dohi, and S. Tsuruta, “Knowledge mining for supporting learning processes,” in *Proc. of the 2008 IEEE Int. Conference on Systems, Man, and Cybernetics (SMC 2008), Singapore*. IEEE, Piscataway, NJ, USA, 2008, IEEE Catalog number CFP08SMC-USB, ISBN 978-1-4244-2384-2, Library of Congress: 2008903109, pp. 2615–2621.

[3] R. Knauf, Y. Sakurai, S. Tsuruta, and K.P. Jantke, “Modeling didactic knowledge by storyboarding,” *Journal of Educational Computing and Research*, vol. 42, no. 4, pp. in press, 2010.

[4] Y. Sakurai, S. Dohi, S. Tsuruta, and R. Knauf, “Modeling academic education processes by dynamic storyboarding,” *Journal of Educational Technology & Society*, vol. 12, no. 2, ISSN 1436-4522 (online) and 1176-3647 (print), pp. 307–333, April 2009.

[5] R. Knauf, Y. Sakurai, S. Tsuruta, K. Takada, and S. Dohi, “Personalized curriculum composition by learner profile driven data mining,” in *Proc. of the 2009 IEEE Int. Conference on Systems, Man, and Cybernetics (SMC 2009), San Antonio, TX, USA*, 2009, ISBN 978-1-4244- 2794-9, pp. 2137–2142.

[6] H. Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, New York: Basic Books, 1993.

[7] R.M. Felder and L.K. Silverman, “Learning and teaching styles in en-gineering education,” *Engineering Education*, vol. 78, no. 7, pp. 647–681, 1988.