An Approach to Automatic Evaluation of Educational Influence

ANI GRUBIŠIĆ, SLAVOMIR STANKOV, BRANKO ŽITKO
Faculty of Natural Sciences, Mathematics and Kinesiology
Nikole Tesle 12, 21000 Split
CROATIA

Abstract: - As the acquisition of knowledge is often an expensive and time-consuming process, it is important to know whether it actually improves the student performance. The e-learning is a revolutionary paradigm that has lately been significantly evolving and it is closely related to the intelligent tutoring systems. Methodology for evaluating the educational influence of learning and teaching process, questions whether and in what amount, students learn effectively. Within this context, we measure educational influence by using the effect size as metric. In this paper we present architecture and functionality of a model for the automatic evaluation of educational influence. The model relies its evaluation engine on methodology that has been originally developed for the intelligent tutoring systems effectiveness evaluation, and it will, generally, be applicable on any e-learning system. That would enable calculation of overall e-learning systems effectiveness, as well as, effectiveness of different categories of e-learning systems, using meta-analysis.

Key-Words: - E-learning, evaluation, efficiency, educational influence, effect size, experiment

1 Introduction

The e-learning presents intersection between a world of information and communication technology and a world of education [17]. When compared with traditional class-room education that centers on teachers who have control over class, educational content and learning and teaching process, the e-learning offers a learner-centered, self-paced interactive, easy-to-access, flexible and distributable learning environment [11]. The most commonly used definition of e-learning is that e-learning is a wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms, and digital collaboration, that make educational content available on different electronic media (CD-ROM, Internet, intranet, extranet, audio and video-tape, satellite, etc.) [2].

The e-learning systems, therefore, provide access to electronically based learning resources anywhere at anytime for anyone [1]. The intelligent e-learning systems have capability to act appropriately in uncertain situations that appear in learning and teaching process. Special class of intelligent e-learning systems are the intelligent tutoring systems (ITS).

Evaluation is useful for the investigation and exploration of different and innovative ways in which technologies are being used to support learning and teaching process. All instructional software should be evaluated before being used in educational process. A useful definition of evaluation could be that evaluation offers information to make decision about the product or process [14]. A well-designed evaluation should provide the evidence, if a specific approach has been successful and of potential value to the others [7]. Each methodology represents a different approach to the evaluation. Which one will be most appropriate, depends on the type of the asked questions. A unique model for the evaluation of e-learning systems' educational influence is hard to define.

Since an overall evaluation of e-learning system's effectiveness is important, effective Web-based assessment tools for the e-learning systems evaluation should be developed [20]. There are numerous applications for creating, delivering and reporting online tests, that is, for the test management (creating, deploying and scoring). We have to emphasize that the test management is required, but not sufficient, for the evaluation of e-learning systems' educational influence. As a result, a need for creating a tool for automatic evaluation of educational influence has emerged. The accent is on the "automatic", because effectiveness can be calculated manually by using different statistical software and different evaluation methodologies, what can be complex and tedious task. A proposed tool for automatic evaluation of educational influence would enable easy effectiveness calculation, as well as, foundation for conducting meta-analysis, that requires effect sizes calculated by using comparable evaluation methodologies.

In this paper, we present the EVEDIN (EValuation of EDucational INfluence), a system for the automatic evaluation of e-learning systems' educational influence. The EVEDIN calculates effectiveness by using information about students' achievements since the students' knowledge is captured by the different online

tests: pre-test, check-point-test and post-test. The test results undergo EVEDIN's statistical analysis mechanism and, as an outcome, the EVEDIN presents the effect of e-learning system's educational influence.

In the second chapter we review the age-long research and development of the Tutor–Expert System (TEx-Sys) [18], a hypermedial authoring shell model for building ITS. In the third chapter we give an overview of evaluation methodology that has been implemented in the EVEDIN. In the fourth chapter we describe EVEDIN's architecture and functionality. Finally, in the last chapter we present the results of using EVEDIN for the automatic evaluation of educational influence of the xTEx-Sys's (eXtended Tutor-Expert System) [19], as the representative of Web-based authoring shells for building ITS.

2 Background

Intelligent tutoring systems are computer systems that support and improve learning and teaching process in certain domain knowledge, respecting the individuality of learner as in traditional "one-to-one" tutoring. The major problems when developing ITS are their expensive and time consuming development process. In order to overcome those problems another approach has been chosen, namely to create particular ITSs from flexible shells acting as program generators.

The first implementation of an intelligent authoring shell model called the TEx-Sys is the on-site TEx-Sys (1992-2001), after that followed the Web-based intelligent authoring shell (1999-2003, Distributed Tutor-Expert System, DTEx-Sys) [15] and, finally, the system based on Web services (2003-2005, xTEx-Sys).

Within the TEx-Sys model, knowledge is represented by semantic networks with frames, which basic elements are nodes and links. Nodes are used for the presentation of domain knowledge objects, while links show relations between the objects. Beside nodes and links, the system supports properties and frames (that are consisted of attributes and their respective values), along with property inheritance. Nodes can also have the following structure attributes: pictures, animations, slides, URL addresses and hypertextual descriptions.

The xTEx-Sys is a Web-based authoring shell with an environment that can be used by the following actors: an expert who designs the domain knowledge base, a teacher who designs courseware for student learning and teaching process as well as tests for the student knowledge evaluation, a student who selects course and navigates trough the domain knowledge content using didactically prepared course content and, finally, an administrator who supervises the system.

Simplified version of evaluation methodology, which is described in the next paragraph, has been used to evaluate the educational influence of the DTEx-Sys. The educational evaluation of the DTEx-Sys had shown that its learning strategy is effective [16]. Namely, the experimental group had shown better results than the control group in every statistical test. Significant statistical difference between the control and the experimental group had revealed some advantages of learning and teaching in the DTEx-Sys over traditional learning. Statistically insignificant difference between the experimental and the control group had revealed capability of the DTEx-Sys to be-come substitution of human tutor. The DTEx-Sys effect size of 0.82 is slightly less than 0.84, a standard value for the intelligent tutoring systems [8].

Results gained through the conducted experiment have shown a need to add some extended functions for courseware development and learning management in the xTEx-Sys in order to get it as close as possible to the Bloom's 2-sigma target [4]. The results from the application of a new, more complex, effectiveness evaluation methodology have been used for determining the xTEx-Sys's effect size.

3 Evaluation Methodology

A well-designed evaluation incorporates a mix of techniques to build up a coherent picture. As an evaluation answers the questions for which it was designed for, therefore the first step in research design is the identification of a research question, for example: "What is the educational influence of an e-learning system on students?" [6]. Hypotheses can be formed after identifying the research question, and must be testable and possible to be confirmed or denied on the basis of specific conditions and results.

The way in which you select your student sample will have both an effect on the gathered information and on the impact that your findings might have. If you pick your own sample of students, you have the opportunity to select the students who are likely to be most cooperative, or a group of students with the most appropriate skill levels. Therefore, a random sample of students should be selected in order to try and get a more representative sample [9].

Different evaluation methods are suitable for different purposes and the development of evaluation is a complex process. Experimental research is common in psychology and education [12] and it is suited for the e-learning system because it enables researchers to examine relationships between teaching interferences and the students' teaching results, and to obtain quantitative measures of the significance of such relationships. Controlled experiment is a way of finding out which aspects of deployed educational intervention are influencing the considered outcomes.

In a variety of different experimental designs [10], we have decided to describe the model and the usage of *pre-and-post test control group experimental design with checkpoint tests* that enable determining the educational effect of the evaluated system. We have added arbitrary number of checkpoint-tests to determine the effectiveness in intermediate states.

Pre and post tests, as well as, checkpoint tests are used because we know that students with different skills and backgrounds come to study a particular subject. We need to establish a base measure of their knowledge and understanding of certain domain knowledge in order to be able to quantify the extent of any changes. So, students involved, should take the pre-test to determine some individual starting level of knowledge or understanding. At a later point, they should take the exact comparable checkpoint tests and pot-test to determine the extent to which knowledge and understanding has been improved by the educational intervention.

3.1 Process of Evaluation

For the purposes of e-learning system effectiveness evaluation, students that are chosen to be part of experiment have to take a 45-minute pre-test that has to be distributed at a very beginning of a course. Pre-test enables obtaining information about the existence of statistically significant differences between the groups concerning student's foreknowledge and its results have to be scored on a 0-100 scale. Based on the results from that test, participating students have to be randomly and equally divided into two groups: the Control and the Experimental groups (equalization of groups using caliper matching mechanism [3]). The Control group will be involved in the traditional learning and teaching process and the Experimental group will use the e-learning system.

Furthermore, both groups have to take several 45-minute checkpoint-tests, as well as, a 45-minute post-test after the end of the course. A number of checkpoint-tests is defined according to the duration of experiment (we recommend no less than one checkpoint-test approximately every four weeks). The checkpoint and the post-test enable obtaining information about the existence of statistically significant difference between the groups concerning evaluating educational influence of e-learning system and their results have to be scored on a 0-100 scale.

3.2 Analysis of Results

First, it has to be checked whether groups' initial competencies were equivalent before comparing the gains of the groups. That means calculating the means of pre-test score for both groups and their standard error of mean. Now, a null-hypothesis has to be stated for every checkpoint-test and post-test: "There is no significant difference between the Control and the Experimental group".

Next, the gain scores from the pre-test to every checkpoint-test and the post-test for both groups have to be calculated. The means of gains for every test and for both groups, as well as, their standard means of error have to be calculated. Then the t-values of means of gain scores have to be computed to determine if there is a reliable difference between the Control and the Experimental group for every testing point (the checkpoints and at the end of the course). If there is statistically significant difference at every testing point (same or slightly rising), it implies that e-learning system has had a positive effect on the students' understanding of the domain knowledge. In other words, the null-hypothesis is rejected.

Whereas the statistical tests of significance tell us the likelihood that experimental results differ from chance expectations, the effect size measurements tell us the size of experimental effect. The effect size is a standard way to compare the results of two pedagogical experiments. Effect size is positive when the experimental group in the study outperforms the control group, and is negative when the control group is better. Effect sizes of around 0.2 are usually considered to be small, 0.5 to be moderate, and 0.8 to be large in size [5].

For determining group differences in experimental research, the usage of standardized mean difference is recommended [13]. The standardized mean difference is calculated by dividing the difference between the experimental and the control group means of gains by the standard deviation of the control group. The following formula (modified according to [13]) is used for the calculation of this standardized score:

$$\Delta_{chk1} = \frac{\overline{X}_{echk1} - \overline{X}_{cchk1}}{s_{cchk1}}, \Delta_{chk2} = \frac{\overline{X}_{echk2} - \overline{X}_{cchk2}}{s_{cchk2}}, \ldots,$$
$$\Delta_{chk2} = \frac{\overline{X}_{echkn} - \overline{X}_{cchkn}}{s_{cchkn}}, \ \Delta_{end} = \frac{\overline{X}_{eend} - \overline{X}_{cend}}{s_{cend}} \qquad (1)$$

where $\overline{X}_{echk1}\{\overline{X}_{echk2}, \ldots, \overline{X}_{echkn}, \overline{X}_{eend}\}$ = mean of gains for the experimental group; $\overline{X}_{cchk1}\{\overline{X}_{cchk2}, \ldots, \overline{X}_{cchkn}, \overline{X}_{cend}\}$ = mean of gains for the control group; $s_{cchk1}\{s_{cchk2}, \ldots, s_{cchkn}, s_{cend}\}$ = standard deviation of gains for the control group using respectively the gain scores from pre-test to every checkpoint-test and post-test.

The average effect size in our approach is calculated as the arithmetic average of partial effect sizes, calculated using (1). It can be easily summarized in the following formula:

$$\Delta = \frac{\Delta_{chk1} + \Delta_{chk2} + \ldots + \Delta_{chkn} + \Delta_{end}}{n+1} \qquad (2)$$

Effect size can be calculated using different formulas and approaches, and its values can diverge. In our

approach to evaluate the educational influence of an e-learning system, we propose computing the average effect size in order to get a unique effect size that can be used in some meta-analysis studies. In other words, effect sizes can be used to express the results from different studies on a single uniform scale of effectiveness.

4 EVEDIN's Overview and Application

As stated before, our goal was to create a system that would enable automatic evaluation of any e-learning system's educational influence. Previously described evaluation methodology has been implemented and makes the central part of the EVEDIN system.

There are numerous applications for creating, delivering and reporting online tests. We have made a selection of four testing tools and compared their features. The selection of those tools has been done after seeking out online testing software using popular searching engines. A brief description of selected tools and their available features are presented in Table 1. These testing systems can be used with any course and may save hours in exams preparation and correction, they may as well save resources like photocopying and distributing the exams papers, locations of these exams, teachers and assistants, etc.

Observed tools enable only the test management (creating, deploying and scoring) and they cannot evaluate the educational influence of e-learning systems. Moreover, their functionalities make only one part of functionalities of the EVEDIN system. The EVEDIN has different question types for designing reusable questions with multimedia, we can score questions and see scoring results, it enables test preview of HTML test, test scheduling and it gives e-mail feedback to the students

Table 1. Comparison of tools for creating and deploying online tests

	Questionmark Perception	Question Writer	SmartLite WebQuiz XP	Articulate Quizmaker
	www.questionmark.com	www.questionwriter.com	www.smartlite.it	www.articulate.com
Test templates	+	-	-	-
Question types				
single answer	+	-	-	+
multiple answer	+	+	+	+
multiple choice	+	+	+	+
true/false	+	+	+	+
matching	+	+	-	+
sequence	+	-	-	+
hotspot	+	-	-	+
fill-in-the-blank	+	+	+	+
essay style	+	-	+	-
short answers	+	-	-	-
numeric entry	+	-	-	+
Reusing questions	+	-	-	-
Adding multimedia	+	+	+	+
Scoring	+	+	+	+
Storing results	+	-	+	-
Wizard layout	+/-	-	+	-
Preview	+	+	+	+
HTML tests	+	+	+	+
Scheduler	+	-	-	-
e-mail feedback	-	+	+	+
Report types	+	-	-	-

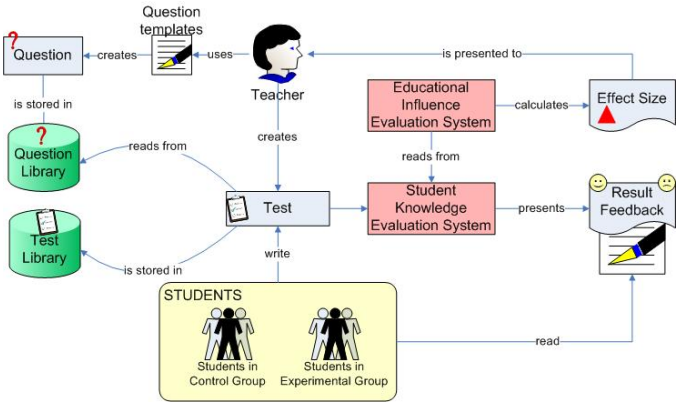

Figure 1. The EVEDIN's functional model

as well as report to the teacher. Besides all that, it enables automatic evaluation of e-learning systems' educational influence.

4.1 EVEDIN'S Functionalities and Architecture

The EVEDIN enables the management and the deployment of experiments, that is, the management of control and experimental groups, creation and deployment of all necessary tests, their automatic scoring and, finally, and what is the most important, it enables calculation of the effect size (see Fig. 1.).

The EVEDIN consists mainly of three independent components: the client, the server, and the database management systems (DBMS). An Internet web server provides access to the HTML pages, the DBMS keeps and classifies all the experiments, users, questions, tests and results in SQL database and ASP.NET applications acts as an interface between the client stations and the DBMS. The client application is divided into two applications, called the teacher and the student applications.

The teacher application is used to organize experiments: it provides users with a tool to manage the control and experimental group (equalization of groups using caliper matching mechanism), questions and tests. The teachers can:

- create, update or delete experiments;
- add, update or delete users and manage control and experimental groups;
- create, update and delete groups of questions;
- create, update or delete questions and answers using templates (single answer, multiple choices, multiple answers, true or false, matching, sequence, fill in the blanks and parallel list connections), store them into questions library and organize them into groups of questions;
- create, update, or delete tests, store them into tests library and publish them to the Web;
- view test results and experiment results.

The teacher application saves much time in test preparation and correction as it can automatically correct test and provide user with results. The teachers can use this tool to print the scores report for the students who have taken the tests. The most distinctive feature teachers can use is the EVEDIN's possibility to calculate the experiment's effect-size based on previously described methodology.

Students can access the tests by using the student application. Student takes test, which is presented in all-question methodology, if it is available. The test will be over either if the student finishes it and select the submit button, or if a pre-allocated time will be over. The system will calculate the score and display the result. Finally, the student score will be saved in the corresponding database for further analysis and calculation of educational influence.

4.2 Automatic evaluation of the xTEx-Sys's educational influence using the EVEDIN

To access the effectiveness of xTEx-Sys, we have conducted an experiment that started in October 2005 and lasted all the way until the end of January 2006. Subjects involved in experiment were undergraduate students at the Faculty of Chemical Technology (FCT) and the Faculty of Natural Sciences, Mathematics and Kinesiology (FNSMK), both at University of Split, that took a course called "An Introduction to Computer Science". Students, who participated in an experiment, have been divided into the Control group (73 students from FCT) and the Experimental group (102 students from FNSMK) before they had written pre-test. That was the reason why the Control and the Experimental groups have been equalized after finishing experiment.

The Control group was involved in traditional learning and teaching process and the Experimental group used the xTEx-Sys. Both groups have taken a 45-minute pre-test at the very beginning of the course. Also, both groups have taken two 45-minute checkpoint-tests at the end of the 5th week (72 Control group students and 99 Experimental group students) and 10th week (57 Control group students and 94 Experimental group students). Furthermore, 49 Control group students and 87 Experimental group students have taken a 45-minute post-test at the end of the course. For the purposes of experiment, the EVEDIN excluded from both groups those students who have failed to attend at least one test. Finally, there were 40 Control group students and 80 Experimental group students whose results could be taken into consideration.

Obviously, those groups were not numerically equivalent, so the EVEDIN has approached to caliper matching of the groups and it simultaneously checked whether newly defined groups had equivalent initial competencies. When the EVEDIN's internal algorithm had defined two subgroups of the Control and the Experimental group that are numerically and statistically equivalent, it proceeded with calculating of the xTEx-Sys's educational influence. The EVEDIN has calculated all necessary means of gains and standard deviations.

The first checkpoint-test had a small partial effect size of 0,17, the second check-point-test had a moderate partial effect size of -0,49 (negative algebraic sign has showed that Control group had performed better) and finally post-test had a large partial effect size of 0,79. The EVEDIN concluded that xTEx-Sys's educational influence has the average effect size of 0,16 sigma, what is according to [5] small effect size. Results presented by the EVEDIN can bee seen in Fig. 2.

Small average effect size of 0,16 sigma should be observed through a prism of partial effect sizes that, in this case, present a nonlinear outcome. Namely, the last partial effect size related to the post-test, reveals about the same effect size that was derived during the DTEx-Sys's efficiency evaluation, where we have calculated only the final effect size. Observed negative peak in the xTEx-Sys's partial effect size calculation happened due to organization problems that appeared just in time of second check-point-test. Because of that problems related to experiment execution organization, the Experimental group has taken the second checkpoint-test before the Control group, what can explain the negative algebraic sign in second checkpoint-test partial effect size and, consequently, a small average effect size.

5 Conclusion

As we have stated in this paper, all instructional software should be evaluated before being used in educational process. A unique model for evaluation of e-learning systems is hard to define and methodology we have presented in this paper can simplify the problem. Presented methodology for evaluation of e-learning systems effectiveness, bases itself on experimental research with the usage of pre-and-post test control group experimental designs with arbitrary number of

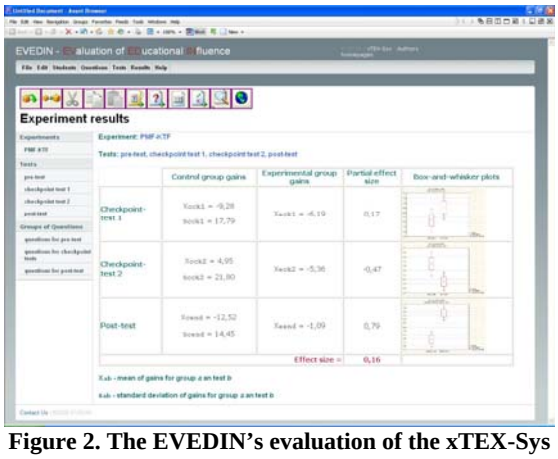

Figure 2. The EVEDIN's evaluation of the xTEX-Sys educational influence feedback

checkpoints.

Our intention was to create a system that would enable automatic evaluation of any e-learning system's educational influence. We have applied the EVEDIN for evaluating educational influence of the xTEx-Sys, a representative of Web-based authoring shells for building ITS, that is a special class of the e-learning systems. The EVEDIN has been found useful for fulfilling this purpose as it enables administration of control and experimental groups, creation and deployment of all necessary tests, their automatic scoring and, finally, what is the most important, it enables calculation of the effect size. Our goal is to use the EVEDIN for evaluating educational influence of the e-learning systems in general. That would enable calculation of the overall e-learning systems effectiveness, as well as, the effectiveness of different categories of e-learning systems using meta-analysis.

6 Acknowledgments

This work has been carried out within projects 0177110 Computational and didactical aspects of intelligent authoring tools in education and TP-02/0177-01 Web oriented intelligent hypermedial authoring shell, both funded by the Ministry of Science and Technology of the Republic of Croatia.

References:

[1] Albert, D.: *E-learning Future – The Contribution of Psychology*. In: Roth, R., Lowenstein, L., Trent, D. (eds.): Catching the Future: Women and Men in Global Psychology, Proceedings of the 59th Annual Convention, International Council of Psychologists, Winchester, England (2001) 30-53

[2] ASTD: *A Vision of E-Learning for America's Workforce*. Report of the Commission on Technology and Adult Learning (2001)

[3] Becker, L.A.: *Online syllabus - Basic and Applied Research Methods*. web.uccs.edu/lbecker/Psy590 /default.html (2000)

[4] Bloom, B.S.: *The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring.* Educational Researcher, 13, (1984) 4-16

[5] Cohen, J.: *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Ass. (1988)

[6] Cook, J.: *Evaluating Learning Technology Resources*. LTSN Generic Centre, University of Bristol (2002)

[7] Dempster, J.: *Evaluating e-learning developments: An overview*, warwick.ac.uk/go/cap/resources/ eguides (2004)

[8] Fletcher, J.D.: *Evidence for Learning from Technology-Assisted Instruction*. In: Technology applications in education: a learning view, (H.F. O'Neal, R.S. Perez (Ed.)), Mahwah, NJ: Lawrence Erlbaum Associates, (2003) 79-99

[9] Harvey, J. (ed.): *Evaluation Cookbook.* Learning Technology Dissemination Initiative, Institute for Computer Based Learning, Edinburgh (1998)

[10] Iqbal, A., Oppermann, R., Patel, A., Kinshuk. A *Classification of Evaluation Methods for Intelligent Tutoring Systems.*, In: Arend, U., Eberleh, E., Pitschke, K. (eds.): Software Ergonomie '99 , B. G. Teubner, Stuttgart, Leipzig (1999) 169-181

[11] Khan, B. H.: *A framework for Web-based learning.* In: Khan, B. H. (ed.): Web-based training. Englewood Cliffs, NJ: Educational Technology Publications (2001)

[12] Mark, M.A., Greer, J.E.: *Evaluation methodologies for intelligent tutoring systems*. Journal of Artificial Intelligence and Education, 4 (2/3) (1993) 129-153

[13] Mohammad, N.Y.: *Meta-analysis of the effectiveness of computer-assisted instruction in technical education and training.* Virginia Polytechnic Institute and State University, Blacksburg, Virginia, PhD Thesis (1998)

[14] Phillips, R., Gilding, T.: *Approaches to evaluating the effect of ICT on student learning*, ALT Starter Guide 8, (2003)

[15] Rosić, M.: *Establishing of Distance Education Systems within the Information Infrastructure.* Faculty of Electrical Engineering and Computing, Zagreb, Croatia, MS Thesis (2000) (in Croatian)

[16] Stankov, S., Glavinić V., Grubišić A.: *What is our effect size: Evaluating the Educational Influence of a Web-Based Intelligent Authoring Shell?* In: Nedevschi, S., Rudas, I. J. (eds.): 8th International Conference on Intelligent Engineering Systems. Cluj-Napoca, (2004) 545-550

[17] Stankov, S., Grubišić, A., Žitko B.: *E-learning paradigm & Intelligent tutoring systems.* In: Kniewald, Z. (ed.): Annual 2004 of the Croatian Academy of Engineering. Croatian Academy of Engineering, Zagreb (2004) 21-31

[18] Stankov, S.: *Isomorphic Model of the System as the Basis of Teaching Control Principles in an Intelligent Tutoring System.* Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Croatia, PhD Thesis (1997) (in Croatian)

[19] Stankov, S.: Project TP-02/0177-01 Web oriented intelligent hypermedial authoring shell, Ministry of Science and Technology, Croatia (2003)

[20] Zhang, D., Nunamaker, J.: *Powering E-Learning In the New Millennium: An Overview of E-Learning and Enabling Technology.* Information Systems Frontier. 5(2) (2003) 207-218