

# ***DERIVING MANAGERIAL KNOWLEDGE FROM AN EDUCATIONAL WEB-BASED INFORMATION SYSTEM***

**Addisson Salazar, Nancy Vargas, Jorge Gosálbez , Ignacio Bosch**

Universidad Politécnica de Valencia

## ***Abstract***

This paper presents the application of a procedure to derive managerial knowledge from an educational web-based information system. The procedure applies a knowledge discovery approach for mining historical databases of the e-learning activities of students at a virtual campus. The web activities include events as course access, email exchange, forum participation, news reading and chats. The techniques applied were principal component analysis, independent component analysis, unsupervised clustering and generation of decision rules. Learning styles of the students were detected and knowledge of global and specific content was found. Those findings were well evaluated by academic experts and they were used to propose some preliminary strategies toward improve academic management and the e-learning system.

*Keywords: e-learning, knowledge discovery, learning style, academic management*

## **1 INTRODUCTION**

Nowadays there exist significant strategies of information systems that use Internet as technological platform for their operations. Business to business (B2B), business to customer (B2C), business to government (B2G), and citizen to government (C2G) are some of the implementations of web-based information systems (White, Daniel, Ward, & Wilson, 2007), (Garrity, Glassbert, Kim, Sanders, & Shin, 2005), (Éthier, Hadaya, Talbot, & Cadieux, 2006), (Paliwal, Adam, Atluri, & Yesha, 2003), (Hof & Reichstädter, 2004). In the education field the web-based information systems have made possible the e-learning in virtual universities. The e-learning activities of the students using the virtual campus information system generate a lot of information that can be exploited in a knowledge discovery process.

This paper presents a case study on knowledge discovery research carried out on data of graduate and undergraduate courses at the Universidad Politécnica Abierta (UPA) site and the proposal of some preliminary strategies derived from the knowledge found. The UPA is a virtual campus at Universidad Politécnica de Valencia and currently it has more than 6000 students registered in about 230 courses, Figure 1 shows a general schema of the virtual campus learning environment. The study pursued to find knowledge about academic performance success and failure of the students and analyzing the e-learning event activity at the campus web to recognize patterns on learning styles of the students. Events covered the personal and collaborative use of the web resources in course activities, including such as content consulting, email exchange, forum participation, etc. The underlying hypothesis was:

there is useful hidden knowledge in data from e-learning web activities for academic management or evaluation of the e-learning system.

Data from the use of the UPA web facilities included the following information about e-learning event activity: course access, agenda using, news reading, content consulting, email exchange, chats, workgroup document, exercise practice, course achievement, and forum participation. Date and time for each event also were available. Besides of the information on the web activity, the exercises achieved and grades obtained by the UPA's students were tried in the knowledge discovery process. The data were collected from the virtual campus web in the period from January 2002 to March 2005, totalizing 2'391,003 records.

The process of knowledge discovery covered the following stages:

- Building a reliable datawarehouse, by filtering data inconsistencies, solving data heterogeneity problems and processing data.
- Obtaining and interpreting patterns of the student behaviour in e-learning activities by using independent component analysis, neural networks and linear regression analysis.
- Obtaining homogeneous data groups by applying clustering processing and selecting data groups, sorted out by research topics, for the definition of decision rules.
- Applying a knowledge representation on selected groups using decision trees to obtain the decision rules of the factors that influence on academic achievement success and failure.
- Evaluating knowledge findings by experts, from the point of view of their validity, novelty, and simplicity.
- Outlining strategies for the improvement of e-learning academic processes.

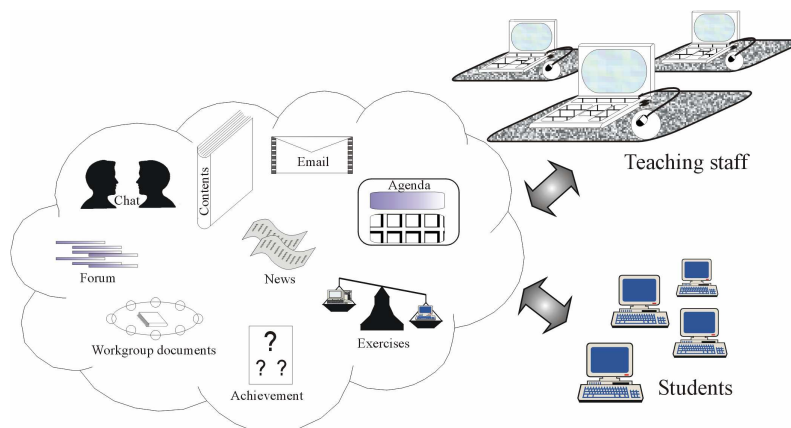


Figure 1. Virtual campus learning environment at UPA..

## 2 BACKGROUND

Knowledge discovery or data mining from web data (webmining) is a new research area that tries to understand the information flow at the web by means of automated techniques for searching knowledge. This area has a wide range of emergent applications including e-learning, e-commerce, automated information assistants and many applications that operate through the web (Srivastava, Cooley, Deshpande, & Tan, 2000). Particularly there is an increasing interest in data mining from web data (webmining) of the e-learning data, some examples are: predicting drop-out on demographic data (sex, age, marital status, etc.) and course data in the first half scores of the course (Kotsiantis, Pierrakeas, & Pintelas, 2003),

predicting the course score processing success rate, success at first try, number of attempts, time spent on the problem, etc. (Minaei, Kashy, Kortemeyer, & Punch, 2003), combining several weak classifiers by boosting to predict final score (Zang & Lin, 2003). Recently, new holistic webmining approaches considering extracting learning styles from the web navigational behaviour of the students have been outlined (Mor & Minguillón, 2004), (Xenos, 2004), (Garcia, Amandi, Schiaffino, & Campo, 2005).

A learning-style model classifies students according to where they fit in a number of scales corresponding to the ways in which they receive and process information. One of the most accepted learning style taxonomy for engineering students is (Felder & Silverman, 1988), see Table 1 (one learning style is conformed by the combination of one feature in each dimension, for instance, intuitive-visual-deductive-active-global). This model was used in the present research.

Preferred Learning Style		Corresponding Teaching Style	
Sensory-Intuitive	Perception	Concrete-Abstract	Content
Visual-Auditory	Input	Visual-Verbal	Presentation
Inductive-Deductive	Organization	Inductive-Deductive	Organization
Active-Reflective	Processing	Active-Passive	Student participation
Sequential-Global	Understanding	Sequential-Global	Perspective

Table 1. Dimensions of Learning and Teaching Styles (Felder et al., 1988) pp. 675.

### 3 DATA PREPROCESSING

A reliable datawarehouse was created from the historical (2001-2005) web data of the UPA. Figure 2 shows a simplified scheme of the data entities at the UPA. The total number of events in the analyzed period was 2'391.003, the projection of the event table on the student code was a 8909 records table. For each student, the corresponding total instance counter of each kind of event was calculated, and a normalized value (1-100 scale) of student event activity was calculated with the following equation,

$$even\_activity_{student} = \frac{event\ total\ instance_{student}}{instance\ maximum_{event}} \cdot 100$$

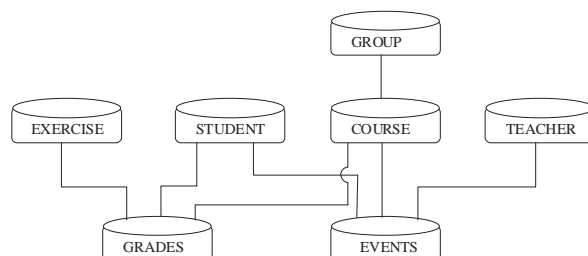


Figure 2. Structure of the web data.

The student activity data were added as fields to the datawarehouse. From the grades table, the average grade for a student was calculated and added to the datawarehouse. Because all the virtual courses do not make evaluation, only 1873 of the rows of the datawarehouse had a value for the variable average grade. To obtain qualitative descriptions of the student event activity, various tables were defined to obtain descriptions of the event activities of the students, see Table 2. Depending on the event activity value for student, the corresponding qualitative event activity description was assigned.

Type	Content consulting
1	Almost never
2	Occasionally
3	Usually
4	Very frequently

Type	Chats
1	Not very active
2	Fairly active
3	Very active

Type	Agenda using
1	Does not use it or use it a little
2	Average use
3	Use it a lot

Table 2. Qualitative description of the student event activity.

#### 4 INDEPENDENT COMPONENT ANALYSIS (ICA)

ICA is a powerful statistical technique that has had a successful application in different areas of signal processing (Hyvärinen, 2001). ICA assumes that there is an  $M$ -dimensional zero-mean vector  $\mathbf{s}(t) = [s_1(t), \dots, s_M(t)]^T$ , such that the components  $s_i(t)$  are mutually independent. The vector  $\mathbf{s}(t)$  corresponds to  $M$  independent scalar-valued source signals  $s_i(t)$ . The multivariate probability density function (p.d.f.) of the vector can be rewritten as the product of marginal independent distributions  $p(\mathbf{s}) = \prod_{i=1}^M p_i(s_i)$ . A data vector  $\mathbf{x}(t) = [x_1(t) \dots x_N(t)]^T$  is observed at each time point  $t$ , such that  $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$  where  $\mathbf{A}$  is called mixture matrix and it is full rank  $N \times M$ , (Hyvärinen, 2001).

ICA was applied on the UPA data in order to identify independent “sources” (independent event activity), i.e. searching those event activity that can separate by an ICA algorithm as a source. After analyzing the results from ICA applied to the different data subsets and considering additional information about the courses and students in the campus, we can infer the following conclusions:

- Email exchange was independent in some cases. It could be due to weakness in teaching strategies for promoting the student interactivity. Then email exchange is transformed in email review done as a routine.
- In courses with no grades, the workgroup document event was independent. The lack of evaluation and grades discourage the participation of students in collaborative tasks.
- In some datasets the content consulting event was independent as reflect of the distributed passive learning (DPL) nature of the web platform. Thus content consulting becomes a routine consisting in download materials with no interactive learning process.
- Exercise practice and course achievement also were found as independent events for some datasets. It could be due the profile of some students that includes information and telecommunications background and knowledge about course contents. For those students participating in those event activities would be irrelevant.

## 5 PRINCIPAL COMPONENT ANALYSIS (PCA) AND ICA

PCA is a very well known technique that reduces the variable order in statistical multivariate analysis (Hardle & Simar, 2006). We applied PCA for grouping the events of the web activity in learning dimensions taking into account the Felder's framework (Felder et al., 1988). PCA reduced 12 variables (10 event activity, average grade and connection time) to 6 components or factors. After analyzing the results of the different outputs of PCA, we found the following relationship between learning style dimensions and event activity variables, see Table 3.

Learning Style		Web event activity
Sensory-Intuitive	Perception	chats, forum participation, course access
Visual-Auditory	Input	news reading, forum participation, email exchange, chats
Inductive-Deductive	Organization	agenda using, exercise practice
Active-Reflective	Processing	course access, workgroup document, content consulting, exercise practice
Sequential-Global	Understanding	course achievement, forum participation, average grade

*Table 3. Dimensions of Felder's Teaching Styles associated with web event activities.*

Note that some web events are associated with more than one dimension; it has sense because a web activity could demand several capabilities of the students used in their learning process. Allowing that kind of relationship we can obtain more real and versatile descriptions of the student learning styles, besides of including all the dimensions of the learning framework. In (Garcia et al., 2005) just three dimensions of the Felder's model were considered and the Bayesian network proposed constrained relationship of the web events with just one dimension of the learning model.

We applied ICA to the 6 components obtained by PCA, Figures 3 and 4 show sources obtained by ICA, respectively, for the grade and no grade graduate course datasets. Figure 3 displays three labelled characterised zones in the learning style space: 1.) Represents the learning style more important in the population. The learning for the students in this zone emphasizes global understanding, active processing, and deductive logic (natural human teaching style), and high grades. 2.) This learning style is focused on inductive logic (natural human learning style), with sequential understanding, and relative active processing. Students within this style could have natural skills for virtual education. 3.) It is characterised by global understanding, deductive logic, and reflective processing. Students within this style would have higher abstraction skills that need of teaching.

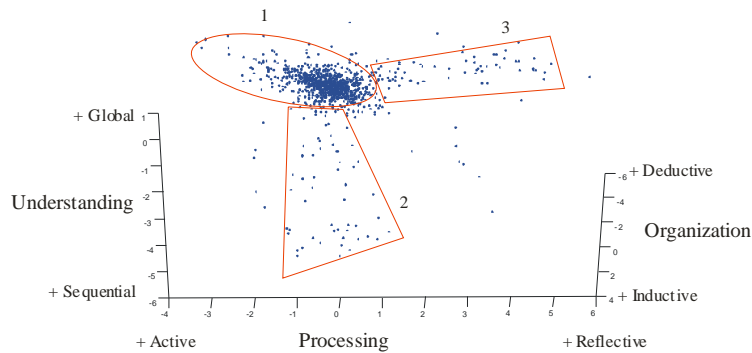


Figure 3. Sources 1-3 in a learning style space for graduate courses with grades.

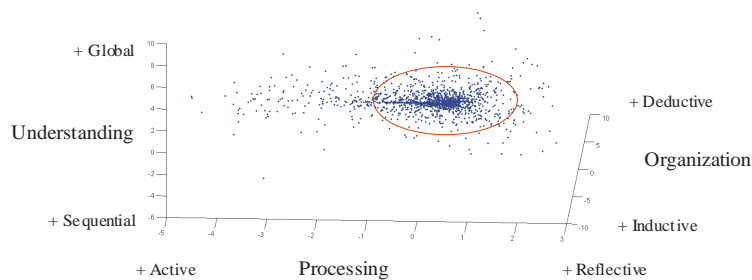


Figure 4. Sources 1-3 in a learning style space for graduate courses without grades.

We can conclude that dimension of understanding enables to project clearly the learning styles, and its principal components are assessment and grades. This finding confirms the assumption that the more quickly way to change the learning style of the student is to change the assessment style, i.e., expected evaluation bias how the student learns (Elton & Laurillard, 1979). Conversely, Figure 4 does not allow forming learning style groups and show all the subjects within a unique learning style. As understanding and organization dimensions do not discriminate projection of the learning styles, only the dimension of the processing provides some discrimination. Then the unique learning style emphasise reflection over actuations, it would be the content consulting and exercise practice components of that dimension. The conclusion is the lack of assessment does not allow developing student learning styles.

## 6 CLUSTERING ANALYSIS

The fuzzy c-means was applied using a fuzziness degree of 1.3 (exponent  $m$ ). Validity measure was the partition coefficient and the maximal number of classes used in training was 16. The calculation was carried out for all classes and the best number of classes was determined and validated by checking the evolution through the class number range of the partition coefficient ( $pc$ ) vs. the classification entropy ( $pe$ ). The application of the fuzzy c-means algorithm on the data subsets generated 23 clusters or groups. Besides of the fuzzy c-means, the conjunctive conceptual algorithm was used on the qualitative variables to get logical conjunctions of relations between the variables. From the analysis of the obtained cluster centroids, the following conclusions were derived.

For the regular career courses with grades:

- The student group with the best grades shows a similar activity level in the different event types, except in course achievement where it has a higher activity than the other groups.
- The student group with the worst grades shows a higher exercise practice proportion than the other groups; however the course achievement is relatively low.
- The intermediate academic performance groups show an imbalance in the proportion of activities, focusing in email exchange and agenda using.
- This data subset does not use events that require interactivity among several students. Those events are chats, forum participation and workgroup documents.
- The students with worst grades are devoted mainly to news reading, content consulting and email exchange but they do not undertake to course achievements.
- The relation cluster-grade follows a normal distribution respecting to the average grade decision variable. Being the best and worst grade groups the less numerous ones.

For the graduate courses with grades:

- There are not significant differences in the academic performance of the clusters.
- The worst grade group is the highest course achievement one.
- The best grade group shows a similar proportion in every event activity, being this group the highest course access.
- In this data subset the events that require interactivity among students were used, but its activity was lower than the events that do not require student interactivity.
- The best grade group was the second most numerous ones and the worst grade group was the less numerous ones.
- The best grade groups used the email more frequently than the others.

For the regular career courses without grades:

- Every group exhibited a good utilization of the interactivity events forum and chats but workgroup documents.
- Every group showed similar proportion in exercise practice and course achievement events.

For the graduate courses with grades:

- Every group showed a high utilization of interactivity events; even the workgroup documents activity was high.
- There was a group with high values for the three interactivity activities, but the value for exercise practice event is very low like in the other groups.

## 7 MINING ASSOCIATION RULES

Average grade was defined as decision variable for mining association rules. 250 decision rules were obtained applying the C4.5 algorithm (Quinlan R.J, 1992) to the clusters of the graduate and regular career course data subsets and to the group consisting of the union of both data subsets. Association rules involve global content or specific content knowledge. Some of the mined association rules are listed below including the success percentage of the rule.

For graduate courses:

Rule No 9: If Agenda using = Average use and  
 News reading = Does not use it or use it a little and  
 Content consulting = Usually and  
 Forum participation = Average Then  
 Average Grade GOOD [80.65%]

Rule No 16: If Content consulting = Very frequently and  
 Course achievement = Usually and  
 Forum participation = Average Then  
 Average Grade GOOD [77.24%]

For the regular career courses:

Rule No 115: If Course = Economic and Financial System and  
 Average time = Evening and  
 Email exchange = Usually exchange it and  
 Course achievement = Very frequently Then  
 Average Grade FAIR [73%]

Rule No 116: If Course = Economic and Financial System and  
 Average time = Evening and  
 News reading = Average use it and  
 Content consulting = Usually and  
 Chats = Fairly active Then  
 Average Grade FAIR [81.5%]

The findings of knowledge obtained in the research were evaluated by academic administration experts of the university in aspects such as validity, novelty, and simplicity, obtaining general score of 8.2 points on a scale from 1 to 10.

## 8 STRATEGIC ACTION OUTLINE

The following are some preliminary strategies that were proposed based on the results of the knowledge discovery process.

- To empower collaborative informatics to include practical virtual labs. Some of the technical subjects to be included are: workgroup, workflow, data mining, searchers, multimedia, and customer research management.
- To collaborate with other educational networks or virtual platforms in order to reinforce teaching quality promoting the creation of virtual learning networks.
- To design and to implement a pedagogic course syllabus for students to understand e-learning education. The appropriate utilization of information and communications technologies helps to educate more and better.
- To adapt the roles of counsellors, teaching, supporting and administrative staffs to the classes in the cyberspace.
- To promote the knowledge of the virtual campus in all the university to generate synergic relationship between university people. This can produce positive feedback to the virtual campus as new students and teachers, e-learning project creation, and communications for improvements.
- To propose special events for diffusion of the virtual university as conferences and workshops on-line, in order to obtain participation of the students.
- To make enable a virtual library with bibliographic contents referenced in virtual courses. In addition of the basic modules and annexes of the courses, access to bibliographical electronic resources allowing research activities would be provided.

## 9 CONCLUSIONS AND FUTURE WORK

The proposed methodology applied to a real case with huge historical data obtained promise results in detecting student learning styles in an e-learning environment. All the dimensions of the Felder's learning framework were modelled using an adaptive approach. The versatility of the approach consists in integrated descriptive modelling using several techniques for processing quantitative and qualitative data. Independent component analysis, a technique normally used in signal processing, has been useful for detecting patterns in e-learning data. Despite of the possible problems of discretization, improvement of interpretation capabilities has been shown. Modelling learning dimensions as a combination of web event activities enhanced the detection of the student learning styles.

The knowledge discovery from e-learning webdata found useful knowledge (of global or particular content) on academic performance of the students at the Universidad Politécnica Abierta (UPA). Among the findings are the following: i.) Events of synchronous interactivity, such as, chats and forum participation and events of asynchronous interactivity empower the student academic performance; ii.) In the courses with grades, academic student performance could be improved by motivating students to have course achievement. Some students show good values for the different event activities, even on exercise practice, but do not have evaluations.

General results of the research were well evaluated by academic experts taking into account the validity, novelty and simplicity of the knowledge. All these knowledge of global and particular contents could be used to improve the e-learning system in different aspects. Strategies to encourage interactivity between students, strategies to design an assessment style to reinforce the student learning style detected, and global improvements of different components of the e-learning system towards a more distributed interactive learning were proposed.

As a prototype, the study has yielded encouraging results on the application of knowledge discovery to e-learning analysis. Nevertheless, in order to obtain a complete application of this analysis is necessary to complement the datawarehouse with more variables. Thus the complexity of the research topic analysis can be more realistically modelled. Student's data among these variables could be, gender, age, location, enrolment date, likes and dislikes, and so on. Besides of teacher's data, such as, course's survey results, research topics, and so on. Those variables would be collected through questionnaires, or transferring automatically from databases.

The teaching part of the Felder's learning framework or another model has to be incorporated in the proposed methodology. The tuning of the learning and teaching styles to obtain a good performance in the outcome of the process would be modelled.

## ACKNOWLEDGEMENTS

Thanks go to the Universidad Politécnica Abierta for giving the web data and information about the virtual campus. This work was supported by Spanish Administration under grant TEC 2005-01820.

## References

- Elton, L. R. B. & Laurillard, D. M. (1979). Trends in research on student learning. *Studies in Higher Education*, 4, 87-102.
- Éthier, J., Hadaya, P., Talbot, J., & Cadieux, J. (2006). B2C web site quality and emotions during online shopping episodes: An empirical study. *Information Management*, 43, 639.
- Felder, R. & Silverman, L. (1988). Learning and teaching styles. *Journal of Engineering Education*, 78, 674-681.
- Garcia, P., Amandi, A., Schiaffino, S., & Campo, M. (2005). Evaluating Bayesian networks' precision for detecting students' learning styles. *Computers & Education*.
- Garrity, E., Glassbert, B., Kim, Y. J., Sanders, G. L., & Shin, S. K. (2005). An experimental investigation of web-based information systems success in the context of electronic commerce. *Decision Support Systems*, 39, 485-503.
- Hardle, W. & Simar, L. (2006). *Applied multivariate statistical analysis*. New York: Springer.
- Hof, S. & Reichstädter, P. (2004). Securing e-Government. *Lecture Notes in Computer Science*, 3183, 336-341.
- Hyvärinen, A. (2001). *Independent Component Analysis*. New York: John Wiley & Sons.
- Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. *Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*.
- Minaei, B., Kashy, D. A., Kortemeyer, G., & Punch, W. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. *Proceedings of 33rd Frontiers in Education Conference*.
- Mor, E. & Minguillón, J. (2004). E-learning personalization based on itineraries and long-term navigational behavior. In *Thirteenth World Web Conference (Ed.)*, (pp. 264-265). New York.
- Paliwal, A. V., Adam, N., Atluri, V., & Yesha, Y. (2003). Electronic negotiation of government contracts through transducers. In *Annual national conference on digital government research dg.o'03*.
- Quinlan R.J (1992). *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann.
- Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: discovery and applications of usage patterns from web data. In *SIGKDD Explorations (Ed.)*, (pp. 12-23).
- White, A., Daniel, E., Ward, J., & Wilson, H. (2007). The adoption of consortium B2B e-marketplaces: An exploratory study. *Strategic Information Systems*, En prensa.
- Xenos, M. (2004). Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education*, 43, 345-359.
- Zang, W. & Lin, F. (2003). Investigation of web-based teaching and learning by boosting algorithms. In *IEEE International Conference on Information Technology: Research and Education (Ed.)*, (pp. 445-449).