Simultaneous evaluation of multiple topics in SIETTE

Eduardo Guzmán & Ricardo Conejo

Departamento de Lenguajes y Ciencias de la Computación.
E. T. S. I. Informática. Universidad de Málaga. Apdo. 4114, 29080 Málaga. SPAIN
e-mail: {guzman, conejo} @lcc.uma.es

Abstract. SIETTE is an efficient web-based implementation of a *Computer Adaptive Test*. The inference machine used is based on *Item Response Theory.* New enhances in the evaluation mechanisms, question selection and finalization criteria have been introduced. New evaluation mechanism allows giving structured knowledge estimation about all topics evaluated in a test. Question selection criteria are able to automatically select a balanced number of items from all topics, so teachers do not need to accomplish this task manually. This paper shows that SIETTE can successfully be integrated into web-based Intelligent Tutoring Systems with structured curriculum, in order to make initial estimations of the student's knowledge level, or even to update the student's model after his exposition to instructional components.

1. Introduction

One of the most important features of an Intelligent Tutoring System (ITS) in comparison to Computer Aided Learning Systems is that ITS adapts to each particular student giving a personalized training [1]. Generally, in web-based ITSs, students can access different instructional components. There students learn different topics, but the ITS does not have any information about how it is accomplished. Therefore, this kind of systems needs some diagnosis module to accomplish a tracking about the examinee's ability. The role of assessment in an ITS is, consequently, the measurement of the state of the knowledge at each time.

Traditionally, ITS have domain knowledge model about the subject being taught, and an instructional planner. The instructional planner decides the best next step in the instructional process. The knowledge model may be composed by a set of nodes that represent the estimated knowledge about each topic of the subject. As a consequence, when a new student begins a course in an ITS, some mechanism is required to have an initial estimation of the student's ability. A first solution could be the realization of a pre-test of each topic involved in the subject. This solution will be acceptable if the number of different topics is small. Otherwise the realization of a test session for each topic can be a very hard and boring task.

SIETTE is a web-based tool to assist teachers and instructors in the evaluation process. Tests offered by this tool are *Computer Adaptive Tests* (CATs) [2], where the evaluation process, the item selection criterion, and the tests finalization criterion are based on a psychometric theory called *Item Response Theory* (IRT) [3].

In this paper, some improvements to the former evaluation mechanisms of SIETTE [4] are presented. The new version of SIETTE is now able to give a structured estimation of the proficiency in each topic after a test session. This estimation mechanism is even able to use multidimensional questions, that is, questions whose resolution depends on the student's knowledge about more than one topic. All these features make the student model of SIETTE more detailed, and as a result, enhance its use as an evaluation module inside web-based ITSs for curriculum-structured domains.

The structure of the paper is as follows: first, a brief introduction to CATs is presented. In section 3, the main components of SIETTE and how the adaptation operates are summarized. Following, the hierarchical structure of the knowledge base of SIETTE is explained. Then the new enhances added to the evaluation mechanism, are introduced, mainly focusing on the *unidimensional* evaluation. Next, an example of a test session is proposed. The paper finishes with some conclusions.

2. Computer Adaptive Tests

A CAT can be defined as a test administered by a computer where the presentation of each item and the decision to finish the test are dynamically adopted based on the student's answers. In more precise terms, a CAT is an iterative algorithm that starts with an initial estimation of the examinee's proficiency level and has the following steps: (1) All the questions in the knowledge base (that have not been administered yet) are examined to determine which is the best item to ask next according to the current estimation of the examinee's knowledge level. (2) The question is asked, and the examinee responds. (3) According to the answer, a new estimation of the proficiency level is computed. (4) Steps 1 to 3 are repeated until the stopping criterion defined is met. This procedure is illustrated in Fig. 1.

The selection and finalization criteria are based upon a Bayesian procedure that can be controlled with parameters that define the required accuracy. The number of questions is not fixed, and each examinee usually takes different sequences of questions, and even different questions.

The main advantage of adaptive testing is that it reduces the number of questions needed to estimate the knowledge level of the student, and that estimation's accuracy is much higher than the estimation achieved by randomly picking the same number of questions [5].

3. The SIETTE system

SIETTE is structured in two parts: (1) a set of author's editors, which allow the teachers to include and modify questions (called *items*), and to hierarchically structure all these *items* in a set of different topics (this structure is called *curriculum*). All these *items pools* are stored in a knowledge base. (2) A virtual classroom where students can take tests about different domains. These tests are generated according to teachers' specifications and are adaptive.

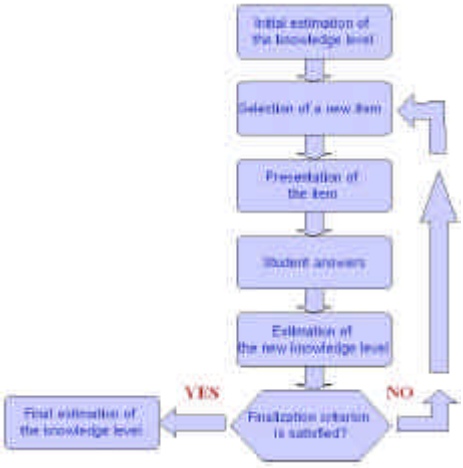

Fig. 1. Flow diagram of an adaptive test. (Adapted from [11])

The system can be used in two different ways: as an independent evaluation tool or as a component of the diagnostic module of an ITS with a curriculum structured knowledge base [6].

While the student is taking the test, the system creates and updates a student model, which mainly stores his knowledge distribution at each stage of the evaluation process. Until this moment, SIETTE has been presented as a system able to measure only one variable called knowledge level (*latent trait* in IRT). This knowledge level is an aggregated value that measures the understanding and know-how of a student in certain subject.

3.1. The adaptation mechanism in SIETTE

In IRT, each item *i* in a test is assigned an *Item Characteristic Curve* (ICC) which is a function, $f : ?? ? , ? ? ?? \; ?0,1?$, representing the probability of a correct answer to that item, given a certain student's knowledge level *?*. Let us represent this probability by the expression: $P(U_i=1| ?)$ or just P_i. Logically, the probability of failing the question is $P(U_i=0| ?) = 1-P(U_i=1| ?)$, or simply Q_i. It is usually assumed that ICCs belong to a family of functions that depend on one, two or three parameters. These functions are based on the normal or the logistic distribution function. SIETTE uses the three-parameter logistic model [7], where the ICC is described by:

$$P_i ? P?U_i ? 1|? ?? c_i ? ?1 ? c_i ?\frac{1}{1 ? e^{?1.7a_i ?? b_i ?}} \quad (1)$$

c_i is the guessing factor, b_i is the difficulty of the question and a_i is the discrimination factor. The guessing factor is the probability that a student with no knowledge at all answers the question correctly. The difficulty represents the knowledge level in which

the student has equal probability to answer or fail the question, besides the guessing factor. The discrimination factor is proportional to the slope of the curve.

The value of *?* is estimated using the response to each item of the test. It is done by a Bayesian method [8], where the probability distribution of the student's knowledge level is calculated, by the Bayes' rule.

A discrete implementation of IRT is used in SIETTE. The latent trait *?* can only take *K* discrete values (from *0* to *K-1*). The ICCs are represented by a vector of *K* components, whose values are initially calculated from the discretizacion of formula (1), but they are dynamically updated by the on-line learning module [5].

If the test is composed by *n* items, given the ICCs, the *a posteriori* estimated knowledge level can be inferred in the following way:

$$\overline{P(? \mid u)} \; ? \left\| \prod_{i?1}^{n} \overline{P_i(?)}^{u_i} (\bar{1} ? \overline{P_i(?)})^{(1? u_i)} \overline{P(?)} \right\| \qquad (2)$$

A distribution of the probability of *?* is obtained applying Bayes' rule *n* times. So, SIETTE does a Bayesian classification of the examinee in one of the *K* classes of knowledge levels according to his answers to the *n* items proposed.

3.2. Hierarchical structure of the curriculum

The items pool is stored into a knowledge base. One for each of the subjects or domains to be evaluated. Each knowledge base is formed by three types of objects:

? *Topics:* They are hierarchically structured forming the *curriculum.* SIETTE can operate with an undefined number of levels in this hierarchy. Each final node of the *curriculum* corresponds to an unique concept or a set of indiscernible concepts in the evaluation sense. Intermediate nodes of the hierarchy represent aggregations of the subtopics of the lower hierarchy according to an inclusion relation. The model of the student associates a knowledge level to each of these (intermediate or terminal) topics. The curriculum structure is defined by the teacher. Independence between nodes of the curriculum that are not directly related is assumed.

? *Items:* They must be explicitly associated to one or more terminal or intermediate topics. This association indicates that the knowledge about a set of topics is required to correctly answer the item. The relation between the knowledge of this topic and the item response is given by an ICC.

In the old version of SIETTE, items only were able to evaluate only one topic. Therefore, each item had one associated *unidimensional* ICC. In the new version, two additional relations can be found between topics and items:

First, items can be associated to more than one topic. For these kind of items, the ICC is *multidimensional* and represents the probability to correctly answer to an item in terms of the combination of the knowledge levels of the required topics. An important constraint that will reduce the complexity of he problems is that an item will be only associated to several topics if these topics are sons of the same topic, i.e., these topics are siblings. In section 4.2 the evaluation mechanism for these items is briefly described.

Second, each item is also associated with all the ancestors in the *curriculum* hierarchy. That is, an item defined to evaluate the topic T_{abc} it can be also used to evaluate the topic T_{ab}, the topic T_a, ; or even the whole subject, where T_{abc} is a subtopic of T_{ab}, etc. This implies having $p+1$ characteristic curves, where p is the depth of the topic to which the topic initially belongs. Let us suppose that each curve is *unidimensional*. These curves represent the probability of correctly answer the item, given the knowledge level of each node respectively. Section 4.1, shows how the evaluation is accomplished in this case.

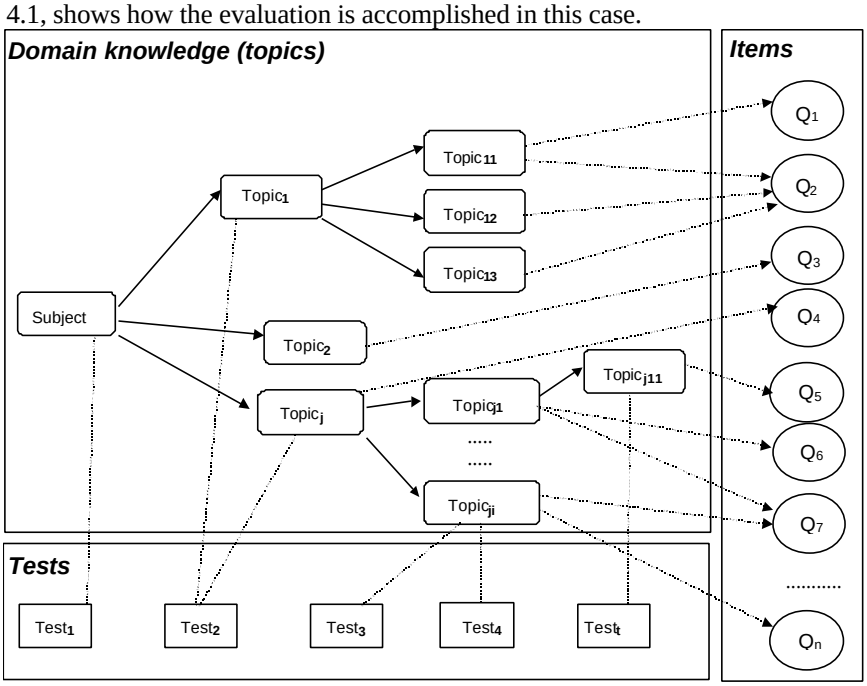

Fig. 2. Structure of the knowledge base

? *Tests:* A test represents an evaluation session. Its main objective is to obtain an estimation of the examinee's knowledge level about one or more topics of the *curriculum.* Therefore, tests are defined in terms of the topic or topics being evaluated. Items which correspond to a test, are those required to accomplish the assessment. There is not a direct relationship between tests and items. This relationship is established through topics. Moreover, in SIETTE, a test can only be associated to sibling topics in the hierarchy.

Two evaluation modes for a test can be prefixed: *aggregated*, if only the evaluation of this node of the *curriculum* is required; or *complete*, if an exhaustive evaluation of all nodes of the sub-tree whose root is the topic, is required.

4. Evaluating multiple topics in SIETTE

In the former version of SIETTE, if a test involved items from several topics, the final knowledge level obtained was a global estimation for all these topics. Also teachers had to explicitly indicate the percentage of items from each topic that appear in a test session.

For an independent evaluation for each topic, a different test for each topic was required. The use of SIETTE as an evaluation module in an ITS requires a more detailed evaluation information. Therefore some extensions of IRT are needed, since classical approach of CAT using IRT as an inference machine are only valid for an aggregated estimation.

4.1. Unidimensional evaluation

The evaluation process is carried out in parallel for each node of the hierarchy, taking the root node topic as the starting point. In this case, the student model is formed by the probability distributions of the knowledge level in each topic assessed. Formally, if a test is composed by the items $Q_1,...,Q_n$ where $(u_1,...,u_n)$ is the vector of responses to these items, the estimation of the knowledge level of the topic k will be obtained from the distribution $P(?_k|u_p,...,u_q)$ which is proportional, like in formula (2), to:

$$? \; k \qquad \mathrm{P}(?_k) \; ? \prod_{i=p}^{q} \mathrm{P}_i(?_k)^{u_i} (1 ? \; P_i(?_k))^{(1 ? u_i)} \tag{3}$$

where $(u_p,...,u_q)$ is the subset of responses to the items associated to topic k or to some of its descendants; $P_i(?_k)$ represents the ICC of the item i, given the knowledge level about the topic k; and $P(?_k)$ is the *a priori* density function or initial estimation of the student knowledge level about the topic k.

For instance, to evaluate the knowledge level of the topic j (T_j), see Fig. 2, all items associated to its descendants can be used. For instance, item Q_5 associated with topic T_{j11}. As a result, the knowledge in topics T_{j11}, T_{j1} and T_j can be simultaneously updated using the characteristic curves associated to item Q_5 for each one of these topics. In the same way, if item Q_n is posed to the student, the knowledge about the topic T_{ji} will be updated. The estimation process of the knowledge level about the topic T_j will be modified too according to the new evidence.

This way of evaluation establishes a particular dependency between the values of knowledge levels of certain topics regarding other topics. Hence, if all items were associated to terminal nodes in the *curriculum*, the process would imply the evaluation of them and the inference of the value of the ascending nodes by the aggregation of their direct descendants. The inverse process is not possible.

Concerning the adaptive mechanism, there are several alternatives. One alternative is to calculate the influence of the possible application of an item in all the student's knowledge level vectors for all the nodes of the hierarchy, and to establish a criterion

of minimum average of the expectations of the *a posteriori* variances. In this case, the system will select an item about the topic whose knowledge estimation is the poorest. As a result, the responsibility of balanced selection of items of each topic is left to the inference machine.

4.2. Multidimensional evaluation

Sometimes, the answer to an item depends on the knowledge of more than one concept (topics in SIETTE). As a result, multidimensional models should be used to evaluate them. In SIETTE, the calculus of the *a posteriori* probability, in this case, is relatively simple. ICCs are transformed into *s* dimensional matrices of *k* components. Clearly, the size of these curves is exponential in terms of the number of topics, but from a practical point of view, the problem is tractable for values of k^s that can be processed in a reasonable response time. The mechanism used to evaluate multidimentional items will be described in future works.

4. An example

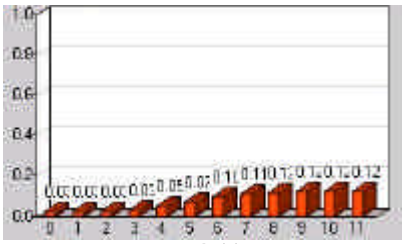

Fig.3 (a)

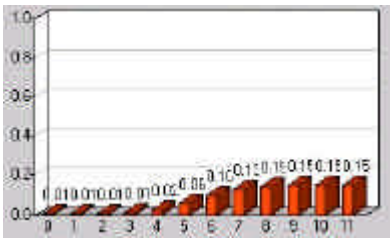

Fig.3 (b)

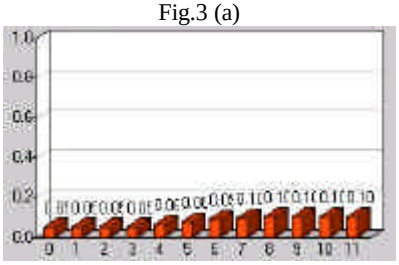

Fig.3 (c)

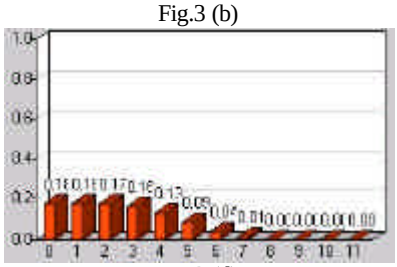

Fig.3 (d)

Fig. 3. Knowledge distribution for each topic after posing five items

In this example, the behaviour of a simulated examinee is going to be analysed. A test session accomplished by the examinee is represented. Let us consider a test of the subject of *Introduction to Compilers* [9]. It is formed by four topics: *(a) Introduction, (b) Lexical Analysis, (c) Syntactic Analysis,* and *(d) Semantic Analysis*. It has been configured with a number of items for each session between 10 and 20. The knowledge is classified into twelve categories (from 0 to 11). All student models

begin with constant knowledge distributions for all topics, equal to 0.083. ICCs of all items are unidimensional with difficulty 5 and discrimination factor 0.7.

The goal of this analysis is to show the enhances in the adaptive mechanism of item selection. These enhances make the system able by itself to automatically choose the most adequate items in order to infer the examinee's knowledge level in each topic. In an item pool where all items have the same difficulty, the adaptation is done according to the topic, not to the item difficulty.

The expected behaviour of the system is that it should pose more items from those topics where the examinee's ability is more uncertainty. That is, if examinee succeeds or fails in the main part of items, the system will be soon able to estimate a knowledge level. In contrast, if he sometimes succeeds but occasionally fails, the estimation process is more difficult, and therefore requires a higher number of items.

Let us suppose that the examinee has good knowledge of topics *(a)* and *(b)*, no knowledge about topic *(d)* and intermediate knowledge about topic *(c)*. He will always answer correctly to items of topics *(a)* and *(b)*, and incorrectly to items of topic *(b)*. For items of topic *(c)*, he will answer to one correctly, incorrectly to the next item, and so on.

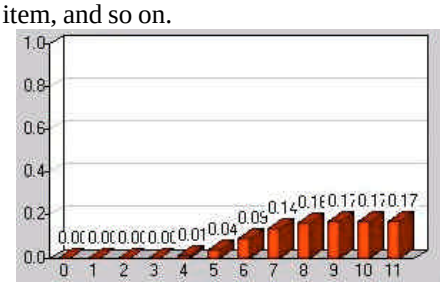

Fig. 4 (a)

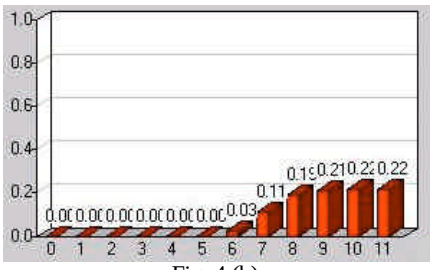

Fig. 4 (b)

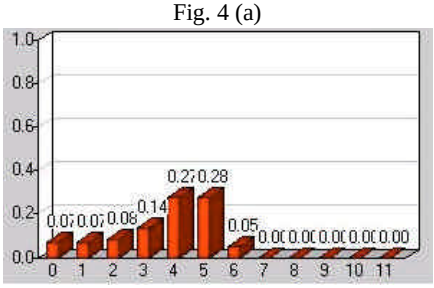

Fig. 4 (c)

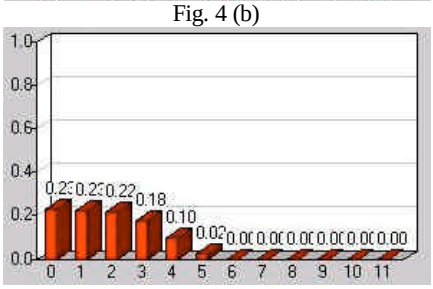

Fig. 4 (d)

Fig. 4. Final knowledge estimation

In a first stage, the system should make an initial estimation of the ability in all topics. In Fig. 3 probability distributions curves of the examinee's knowledge after posing five items are shown. All distributions except the one for topic *(c)* (Fig. 3 (c)) begin to have a certain slope, which indicate that examinee has a well defined level in these topics. On the other hand, the curve for the topic *(c)* is not so inclined because the examinee has given right answers to some items and wrong answers to other items. The test has not reach at a conclusion about his knowledge level in this topic. The estimation of the knowledge is more difficult in this case. Let us take into account that all items have a difficulty equal to 5. Reasonably, examinee should be

classified into the knowledge level 5, because his knowledge is intermediate, but he is failing half of items, which makes the estimation process more complex.

Now the expected behaviour of the system is that to pose more items for topic *(c)* than for the other topics, since the estimation is more difficult. The curves for the other topics will have each time a smaller variance, i.e., its dispersion will be smaller and the estimation more accurate.

The test finished because the maximum number of items has been reached and not because the adaptive finalization criterion (Fig. 4). For this reason, although curves show low variance, any of the levels have a probability significantly greater than the former level.

Table 1 shows the final percentage of items posed to the examinee. The maximum corresponds to topic *(c)* which makes sense because the behaviour of the student was not consistent, so the knowledge level of the topic has been more difficult to estimate. In spite of that fact, table 1 shows that the percentage of items posed from each topic is balanced. Obviously, the knowledge of those topics in which the examinee has always succeeded or failed has been quickly estimated.

Topic	*Correctly answered items*	*Incorrectly answered items*	*Total number of items*	*Estimated knowledge level*	*Percentage of total items*
(a)	*3*	*0*	*3*	*11*	*15%*
(b)	*6*	*0*	*6*	*11*	*30%*
(c)	*5*	*4*	*9*	*5*	*45%*
(d)	*0*	*2*	*2*	*0*	*10%*

Table 1. Statistics of the first examinee's test session

5. Conclusions

An improved adaptive mechanism based on IRT has been presented. SIETTE provides tests where student's knowledge level can be estimated according to the topics, subtopics and concepts in the hierarchically-structured curriculum. These features make this system useful as a diagnostic tool in a web-based ITS. On the other hand, Web based ITSs generally use instructional components which do not give feedback about the influence of tutorial component in the student's learning process. In this case, SIETTE can be integrated in such tutorial components to compensate their lack of feedback mechanisms. By means of a test of all topics covered by the instructional component, the ITS can obtain information about how this instruction has modified the student's proficiency.

In previous versions of SIETTE, teachers were enforced to set the percentage of items to be presented to examinees. This was done to guarantee that items of all topics were presented and that the number of items of each topic was balanced. Thanks to the modifications introduced, the adaptive mechanism of item selection is able to automatically select the most adequate percentage of items of each topic. This is an implicit consequence of the searching for a better estimation of the knowledge in each topic.

The multidimensionality introduced in the curriculum makes for a more realistic assignment of items to topics. Often teachers are enforced to assign an item to a certain topic, when it could be also assigned to other topic. In these cases, the assignment to only one topic despises relevant information for the estimation of knowledge about other topics. The use of multidimensional items has also some influences in the finalization of the test, since these items are useful to estimate knowledge in several topics, and as a result, a lower number of items is required to finish the test.

On the other hand, the main problem of the integration of SIETTE in an web-based ITS is that the ITS and SIETTE must have a common structured *curriculum*, or at least, a correspondence may be established between the student model used in the tutoring system and the *curriculum* settled in SIETTE. Moreover the values returned by SIETTE are coarse data that has been obtained from the observations. ITSs might have mechanisms to indirectly infer the values of knowledge in certain topics from knowledge estimation in other topics. For instance, if a student has demonstrated a high knowledge level in a topic, a low knowledge level in another topic, which is one of its prerequisites, will be very unlikely. SIETTE does not manage this kind of relations between topics. These inferences must be accomplished by the ITS. There are systems [10] that manage more complex models, but it implies a very high computational cost.

The system can be tested at http://www.lcc.uma.es/SIETTE

References

[1] Self, J. (1990) Theoretical foundation for Intelligent Tutoring Systems. AAAI/AI-ED 1990; 45.

[2] Wainer H. (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[3] Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[4] Ríos A., Millán E., Trella M., Pérez-de-la-Cruz J., Conejo R. (1999). *Internet Based Evaluation System*, In: *Artificial Intelligence in Education AIED'99*, Le Mans (1999) 387-394.

[5]Conejo, R., Millán, E., Pérez-de-la-Cruz, J., Trella, M., (2000). An empirical approach to on-line learning in SIETTE. In *Proceedings of ITS'2000*, Montreal. Springer-Verlag. 57-60.

[6] Trella, M., Conejo, R. (2000). ITS Web based Architecture for Hierarchical Declarative Domains, in: *Young Researchers Track Proceedings, ITS'2000*, Montreal 57-60.

[7] Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's mental ability*. In Lord, F. M. & Novick, M.R. (ed.) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

[8] Owen, R. J. (1975). A Bayesian sequential procedures for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association 70*, 351-356.

[9]Aho, A.V., Sethi, R., Ullman, J.D. (1987). Compilers: principles, techniques and tools.

[10]Millán, E., Pérez-de-la-Cruz, J. L., Suárez, E. (2000). An Adaptive Bayesian Network for Multilevel Student Modeling. In *Proceedings of ITS´2000*. 534-543.

[11] Olea J.,Ponsoda, V.: Tests adaptativos informatizados. In Muñiz, J.(ed) *Psicometría*. 1996. Madrid: Universitas.