Modelado del alumno: un enfoque bayesiano

Ricardo Conejo, Eva Millán, José Luis Pérez de la Cruz y Mónica Trella

ETSI Informática, Universidad de Málaga. Apdo 4114, Málaga 29071
{conejo,eva,perez,trella}@lcc.uma.es

Resumen

El diagnóstico es uno de los procesos más importantes dentro de cualquier STI. En este trabajo se propone el uso de la teoría de la probabilidad como marco teórico para el diagnóstico del alumno. Además, se conecta este problema con la teoría de los test adaptativos informatizados.

Palabras clave: Sistemas tutores inteligentes, modelado del alumno

1 Introducción

El modelado del alumno es un problema central en el diseño y desarrollo de un sistema tutor inteligente (STI). En efecto, si la característica que distingue a los STIs de los sistemas de Enseñanza Asistida por ordenador tradicionales es su capacidad de adaptación al alumno [Shute, 95], el sistema debe ser capaz de determinar con la mayor precisión y rapidez posible cuál es su estado cognitivo, es decir, qué partes del dominio que pretendemos enseñarle son las que ya domina y cuáles son las que aún desconoce. Sólo de esta forma será posible adaptar el proceso instructor: saber qué estrategia instructora es más conveniente, qué acción recomendarle (estudio, resolución de ejercicio, juego), qué ejercicio se adecua más a su nivel de conocimiento, etc.

El problema del modelado del alumno puede dividirse en dos componentes: (a) seleccionar la estructura de datos (*modelo del alumno*) que será usada para representar toda la información relativa al alumno: estado cognitivo, estrategias instructoras preferidas, pantallas visitadas, ejercicios resueltos, etc.; y (b) elegir el procedimiento que utilizaremos para realizar el *diagnóstico*, es decir, para inferir dada la información generada en la interacción del alumno con el sistema (problemas resueltos, pantallas visitadas, etc.) el estado cognitivo del alumno. Evidentemente ambas componentes están íntimamente relacionadas, y por tanto lo ideal es diseñarlas y desarrollarlas simultáneamente. Una descripción más completa del problema se puede encontrar en [VanLehn, 88].

El diagnóstico es sin duda uno de los procesos más importantes dentro de cualquier STI, puesto que como ya hemos mencionado, de la calidad del modelo del alumno dependerá la capacidad de adaptación del sistema. Desgraciadamente, no siempre se le presta la atención que merece, dado que el gran esfuerzo que supone desarrollar un sistema tutor inteligente hace que a menudo el problema del modelo del alumno se resuelva mediante la aplicación de heurísticos diseñados a tal fin. Pero la falta de consistencia de dichos heurísticos hace que el comportamiento del sistema sea impredecible, sobre todo en situaciones diferentes a las inicialmente previstas por sus diseñadores. Es por ello por lo que pensamos que, pese al esfuerzo adicional que supone, merece la pena utilizar teorías bien fundamentadas y ampliamente comprobadas que garanticen el funcionamiento óptimo del sistema en todas las situaciones posibles, y, en concreto, proponemos el uso de la *teoría de la probabilidad* como marco teórico. Además, queremos conectar el problema del

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. No.12 (2001), pp. 50-58.
ISSN: 1137-3601. © AEPIA (http://www.aepia.dsic.upv.es/).

modelado del alumno con la teoría de los *test adaptativos informatizados* (TAI) [Wainer, 90], que se ha desarrollado dentro del campo de la psicometría y que pese a su capacidad demostrada para mejorar el proceso de diagnóstico tanto en precisión como en tiempo [Huang, 96] no ha sido aún utilizada dentro del campo de los STIs.

Este artículo se estructura de la siguiente forma: en la siguiente sección describimos los conceptos básicos de los test adaptativos informatizados, así como el sistema SIETTE, que es una herramienta web basada en la teoría de la respuesta al ítem unidimensional que cumple con dos objetivos distintos: (a) permite que los profesores definan de una forma muy sencilla un test adaptativo informatizado; y (b) permite que los alumnos realicen los test definidos y sean evaluados por el sistema, todo ello a través de la web [Ríos, Millán et al., 99]. Posteriormente, describimos un enfoque basado en redes bayesianas [Pearl, 88] que permite realizar test adaptativos en los que se mide más de una habilidad, más adecuados si el objetivo del test no es meramente evaluar al alumno sino llevar a cabo el proceso de diagnóstico en un STI. Finalmente, presentamos las conclusiones obtenidas en la realización de ambos trabajos y las líneas futuras de investigación.

2 Diagnóstico basado en el modelo TRI

2.1 Tests adaptativos informatizados

El uso de los tests para la evaluación es una técnica ampliamente usada en el campo de la educación. Los métodos tradicionales de diseño y administración de tests dependían en gran medida de que éstos fuesen orientados a un individuo o a un grupo. Los tests administrados a grupos son menos costosos en tiempo y recursos que los individuales y además tienen la ventaja de que todos los examinandos están en igualdad de condiciones. Como contrapartida, los tests de este tipo deben contener ítems con tantos niveles de dificultad como posibles niveles de conocimientos puedan existir en el grupo de alumnos que va a realizarlos, mientras que los tests administrados individualmente contienen ítems elegidos de forma más apropiada.

Este hecho puede acarrear consecuencias no deseables como el aburrimiento de alumnos con niveles altos de conocimiento o el desconcierto y la frustración en los alumnos menos aventajados. A principios de los 70 surgieron trabajos que apuntaban que el uso de tests más flexibles aliviaría en parte estos problemas. En [Lord, 70] se establece la estructura teórica de un test de administración masiva pero adaptado individualmente: "*la idea básica de un test adaptativo es imitar lo que un examinador sensato haría*" [Wainer & Messick, 83], es decir, si un examinador hace una pregunta que resulta ser demasiado difícil, la siguiente debería ser más fácil. Sin embargo, probar los tests adaptativos de una forma seria no fue posible hasta principios de los 80, con la aparición de ordenadores potentes y menos costosos. Surgen entonces los llamados test adaptativos informatizados (TAI). Un Test adaptativo informatizado es básicamente un test administrado por ordenador donde la presentación de cada ítem y la decisión de finalizar el test se toman de forma dinámica basándose en la respuesta del alumno y en la estimación de su nivel de conocimiento.

En términos mas precisos, un TAI es un algoritmo iterativo que que comienza con una estimación inicial del nivel de conocimiento del alumno y que tiene los siguientes pasos: (1) Todas las preguntas que no se han administrado todavía son examinadas para determinar cual será la mejor para ser propuesta a continuación, según el nivel de conocimiento estimado del alumno; (2) la pregunta es planteada y el alumno responde; (3) de acuerdo con la respuesta del alumno, se realiza una nueva estimación de su nivel de conocimiento. Los pasos del 1 al 3 se repiten hasta que se cumpla alguno de los criterios de terminación definidos.

Los TAIs tienen importantes ventajas frente a los tests tradicionales a lápiz y papel entre las que destacan: (a) decremento significativo en la longitud de los tests; (b) estimaciones más precisas del nivel de conocimiento del alumno; (c) mejora en la motivación de los alumnos; (d) se puede almacenar un gran banco de preguntas, incluyendo enunciados y posibles respuestas con contenido multimedia.

Los elementos básicos de un TAI son:

- **Modelo de respuesta del ítem**. Este modelo describe como el sujeto responde al ítem según su nivel de conocimiento. Cuando se llevan a cabo mediciones del nivel de conocimiento, cabe esperar que el resultado obtenido no dependa del instrumento utilizado, es decir, la medida ha de ser invariante con respecto al tipo de test y al sujeto al que se le aplica el test.
- **Banco de preguntas**. Constituye uno de los elementos fundamentales para la creación de un TAI. Para definir un banco de preguntas

eficiente se deben especificar las distintas áreas de conocimiento del dominio. Una vez hechas las especificaciones del contenido del test, el banco de preguntas debe contener ítems en suficiente número, variedad y niveles de dificultad [Flaugher, 90].

- **Nivel de conocimiento de entrada**. Elegir de forma adecuada el nivel de dificultad de la primera pregunta que se realice en el test puede reducir sensiblemente la longitud del mismo. Para ello se pueden usar diferentes criterios como tomar el nivel medio de los sujetos que han realizado el test previamente, o crear un perfil de sujeto y usar el nivel medio de los alumnos con un perfil similar [Thissen & Mislevy, 90].
- **Método de selección de preguntas**. Un test adaptativo selecciona el siguiente ítem que va a ser presentado en cada momento en función del nivel estimado del conocimiento del alumno y de las respuestas a los ítems previamente administrados. Seleccionar el mejor ítem puede mejorar la precisión en la estimación del nivel de conocimiento y reducir la longitud del test.
- **Criterio de terminación**. Para decidir cuándo debe finalizar un test se pueden usar diferentes criterios tales como parar cuando se haya alcanzado una precisión determinada en la medida del nivel de conocimiento, cuando se hayan planteado un número determinado de ítems, etc.

2.2 Teoría de respuesta al ítem

La mayor parte de las aplicaciones prácticas de la teoría de la medida en Psicología y Educación están basadas en la Teoría Clásica de Tests (TCT), cuyas deficiencias alentaron la búsqueda de modelos alternativos. Entre los que mayor difusión han tenido destacan los basados en *la Teoría de la respuesta al ítem* (TRI) [Lord, 68], [Hambleton, 89], inicialmente conocida como *teoría del rasgo latente*. La TRI, partiendo de hipótesis restrictivas, intenta dar fundamentos probabilísticos al problema de la medición de rasgos no observables. Su nombre es debido a que se consideran los ítems como las unidades básicas de los tests.

Todos los modelos TRI tienen unas características comunes: (a) suponen la existencia de rasgos o aptitudes latentes que permiten predecir o explicar la conducta de un examinando ante un ítem de un test; (b) la relación entre el rasgo y la respuesta del sujeto al ítem puede describirse por medio de una función monótona creciente, denominada Curva característica del item (CCI).

Los primeros modelos aparecidos son conocidos con el nombre de ``modelos normales'' ya que la forma de la ICC era la de una distribución normal [Lord, 68]. Las dificultades para el manejo analítico de esta función llevaron a los *modelos logísticos*, basados en la función de distribución logística, entre los que destacan los de un parámetro [Rasch, 60], y los de dos y tres parámetros [Birnbaum, 68]. Todos estos modelos están basados en la suposición de independencia local que afirma que si la aptitud θ que explica el rendimiento en el test permanece constante, las respuestas de los examinandos a un par de ítems cualquiera, son estadísticamente independientes. En el modelo de tres parámetros la CCI_i del item i indica la probabilidad de que el alumno responda correctamente ($u_i=1$) supuesto que $\theta = x$ mediante la expresión

$$P_i(x) = P(u_i=1/\theta=x) = c_i + \frac{1-c_i}{1+e^{-1.7ai(x-bi)}}$$

donde a_i es el índice de discriminación, b_i es el grado de dificultad y c_i es el factor de adivinanza.

Este modelo de respuesta se usará para obtener la estimación del rasgo θ. Hay varios métodos para ello. Por ejemplo, en *el método de máxima probabilidad* [Lord, 80] se busca el valor de θ que hace máxima la función de probabilidad:

$$L(\overline{u} / \theta = x) = L(u_1,...,u_n / \theta = x) =$$

$$= \prod_{i=1}^{n} P_i(x)^{u_i} (1-P_i(x))^{1-u_i}$$

donde $\overline{u} = (u_1, ..., u_n)$ es el vector de las respuestas del alumno, es decir, para $i = 1, ..., n$, u_i es 1 si la respuesta al ítem iésimo es correcta y 0 en caso contrario, y $P_i(x)$ es la probabilidad de responder correctamente al ítem i cuando el nivel de conocimiento es $\theta = x$.

Por otra parte, el *método bayesiano* calcula el nivel de conocimiento para el que la distribución a posteriori es máxima. Esta distribución es proporcional al producto de la función de probabilidad y la función de densidad a priori, es decir, $P(\theta/u) \propto L(\theta/u)\, f(\theta)$.

En cuanto a los métodos de selección los más comunes son el de la *máxima información* [Weiss & Kingsbury, 84], que consiste en seleccionar el ítem que haga máxima la información del ítem para el

nivel de conocimiento estimado hasta el momento, y los *bayesianos*, como el de Owen [Owen, 75], que selecciona la pregunta que hace mínima la varianza a posteriori de la distribución del conocimiento.

2.3 El sistema SIETTE

El sistema SIETTE [Ríos, Millán et al., 99] se basa en una implementación discreta de la teoría de test adaptativos, a la que se le han añadido algunos nuevos elementos para mejorar su funcionalidad y para la que se ha utilizando la WWW como interfaz de desarrollo y realización de test.

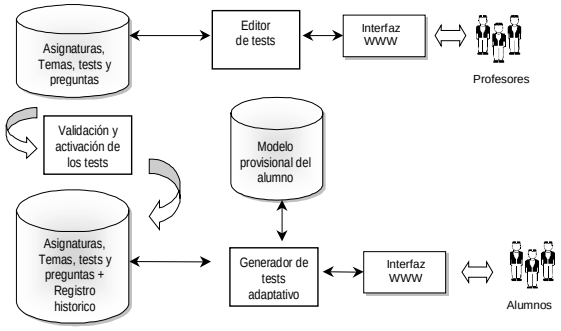

Figura 1. Arquitectura general de SIETTE

La arquitectura general del sistema SIETTE se muestra en la Figura 1 y consta de dos módulos claramente diferenciados. El primer módulo ha sido diseñado para que un conjunto de profesores pueda insertar preguntas y definir los tests a realizar. Además de las preguntas, el profesor puede definir un conjunto de temas en los que se divide la matera, organizar las preguntas de acuerdo a estos temas, definir el numero de respuestas posibles a mostrar al alumno para cada pregunta, así como los parámetros propios de la composición de cada tests, como el porcentaje de preguntas de cada tema que en el intervienen, el modo de selección de preguntas y el criterio de finalización, el numero mínimo y máximo de preguntas a realizar, etc. El módulo de generación de tests es al que accede el alumno para realizar las pruebas, las preguntas se generan de forma individualizada para cada alumno según la materia y el test que haya elegido. Se mantiene un registro temporal de la evolución del alumno durante la sesión que se tiene en cuenta en el proceso de selección de preguntas y un registro histórico de las respuestas a cada pregunta que servirá como fuente de información para el aprendizaje automático de los parámetros de las preguntas. Se ha incluido en el sistema un proceso automático de validación y activación de las preguntas y de los tests diseñados por los profesores que permite, por un lado, separar las tareas de edición y realización de tests, eliminando los posibles problemas de inconsistencia en el caso de edición y realización simultánea; y, por otro lado, se encarga de comprobar que la definición de las preguntas y de los tests es coherente (por ejemplo, si existe un número suficiente de preguntas de cada tema, si se han incluido suficientes respuestas alternativas para cada pregunta, etc.)

La estructura del sistema SIETTE se basa en los conceptos tradicionales de asignatura y tema. SIETTE puede trabajar de forma simultánea con varias materias independientes. El acceso al sistema se efectúa mediante una clave asociada a cada asignatura. Esta clave puede ser personal o compartida por un conjunto de profesores de dicha materia. Cada asignatura o materia se subdivide en temas en la forma que el profesor crea mas adecuada, usando para ello el editor para añadir, modificar o eliminar temas. Los temas representan grandes bloques de la asignatura y no necesariamente conceptos concretos. En su versión actual SIETTE no maneja ninguna información sobre la interdependencia de los temas.

Una vez definidos los temas, el profesor debe definir las preguntas o cuestiones que compondrán los tests. Para ello utiliza el editor que aparece en la Figura 2 y rellena una ficha en la que se incluye el enunciado, la respuesta correcta y una o varias alternativas de respuestas incorrectas. El sistema ha sido diseñado de forma que también almacene una posible ayuda, asociada a cada posible respuesta del alumno, o en caso de que éste requiera mayor información para resolver la pregunta. Las preguntas y las respuestas se pueden introducir como texto simple al que, si así se desea, se puede añadir de código HTML, JavaScript, *applets* y metacódigo PHP. Por consiguiente el formato de las cuestiones admite cualquier objeto multimedia. Además del enunciado, las respuestas y la ayuda, el profesor debe proporcionar algunos datos sobre la cuestión como, por ejemplo, el número de alternativas incorrectas a mostrar, el grado de dificultad y el factor de discriminación de la pregunta, la distribución en pantalla, etc. y debe asociar cada pregunta a uno o varios temas de entre los definidos anteriormente. Dado que el enunciado y las respuestas permiten el uso de metacódigo y applets, cada cuestión introducida puede ser en realidad un esquema generador de cuestiones, lo que permite una gran variedad de preguntas. También es posible simular como quedará la pregunta al presentarla. Esto es especialmente útil en el caso de preguntas generativas.

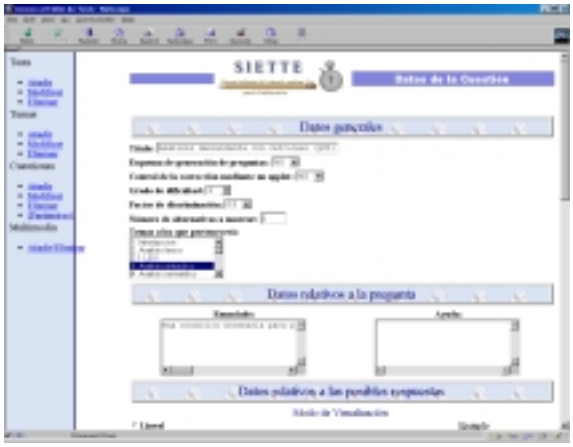

Figura 2. Edición de tests

Una vez definidas las preguntas de una materia y asociadas cada una de ellas a uno o varios temas, el profesor puede definir los tests que se van a realizar. Un test se compone de un conjunto de preguntas seleccionadas según diversos criterios basados en la teoría de tests adaptativos. Al definir el test deben seleccionarse los criterios que se usarán tanto para la selección de preguntas como para la finalización del test. A diferencia de los sistemas clásicos en TRI/TAI, y a fin de ajustar la composición de los tests, el profesor puede especificar los porcentajes de preguntas de cada tema que compondrán cada test, y el número mínimo de preguntas de cada uno de ellos. Esto debe garantizar que un alumno que supere el test tiene un conocimiento suficientemente homogéneo de la materia. También por cuestiones prácticas se fija un número mínimo y máximo de preguntas para garantizar que en cualquier caso el test tiene un final, se alcance o no el criterio de finalización estadístico.

El algoritmo para la generación de tests es el expuesto más arriba: (a) selección de la pregunta: que puede ser aleatoria, adaptativa o por máxima probabilidad, y bayesiana o por mínima desviación tipica a posteriori. También se tienen en cuenta la distribución de temas en la composición de los tests que ha indicado el profesor y aspectos funcionales tales como la no repetición de preguntas en un mismo tests o en sucesivos tests realizados por el mismo alumno; (b) estimación del nivel de conocimiento del alumno; y (c) criterio de finalización: según las especificaciones del profesor y atendiendo a la desviación típica de la variable estimada, la cota del error sobre un determinado nivel y los mínimos y máximos de preguntas previamente configurados.

A diferencia de la teoría clásica de la TRI/TAI, que trabaja con funciones definidas sobre el conjunto de los números reales, SIETTE emplea una aproximación numérica a estas funciones. Esto se traduce en una mayor facilidad para la aplicación de los métodos bayesianos ya que no es necesaria la resolución exacta de las ecuaciones. Igualmente desde el punto de vista computacional el proceso es mas eficiente si el número de intervalos considerados es pequeño. Por otra parte, el uso de aproximaciones numéricas tiene la ventaja de que no está restringido a ninguna familia de funciones concreta para definir las ICC de las cuestiones. Si bien se ha seguido utilizando la función logística como base para la definición de estas curvas de forma paramétrica por parte del profesor, nada impide utilizar otras distribuciones que no se ajusten a esta familia de curvas, con lo que el ajuste a la distribución real de dificultad de las preguntas puede ser mejor. Este ajuste puede llevarse a cabo mediante técnicas de aprendizaje estadístico de forma directa [Conejo, Millán et al., 2000], sin necesidad de estimaciones restringidas a una familia concreta de curvas.

SIETTE permite el acceso tanto de forma identificada como anónima. En el caso de usuarios registrados, el sistema tiene en cuenta los tests realizados anteriormente por el alumno, y es capaz de mantener el modelo temporal del alumno como punto de partida para una nueva evaluación. El alumno puede seleccionar el test a realizar de entre todos los test disponibles y puede configurar ciertos parámetros, como la presentación de las respuestas correctas inmediatamente después de la resolución de cada pregunta o sólo al final del test.

3 Diagnóstico mediante redes bayesianas

En esta sección vamos a describir cómo las redes bayesianas pueden ser utilizadas en el problema de diagnóstico del alumno. Para ello definimos en primer lugar el modelo estructural que servirá como soporte del proceso evaluador (nodos, enlaces y parámetros), y después presentamos los resultados obtenidos en la evaluación del modelo propuesto. Dicha evaluación ha sido realizada utilizando alumnos simulados.

3.1 Modelo estructural

Para utilizar redes bayesianas en el problema de diagnóstico, lo primero que tenemos que hacer es definir los elementos básicos: variables, enlaces entre ellas y parámetros. A continuación presentamos el modelo estructural integrado que

hemos desarrollado, que no sólo permite realizar el diagnóstico a diferentes niveles de granularidad, sino que propone simplificaciones notables para la especificación de los parámetros. Nos centraremos en el diagnóstico basado en preguntas tipo test, aunque en principio sería posible considerar cualquier tipo de preguntas siempre que el sistema tuviera la capacidad de comprobar si la solución propuesta por el alumno es o no correcta.

3.1.1 Variables

Consideraremos dos tipos básicos de variables: variables para medir el grado de conocimiento alcanzado por el alumno, y variables para recolectar evidencia. A su vez, y para una evaluación más detallada, las variables de conocimiento se definen a diferentes niveles de granularidad. Describimos a continuación cada uno de estos tipos.

Variables para medir el conocimiento del alumno. Vamos a utilizar tres niveles de granularidad, que creemos que serán suficientes en la mayoría de las aplicaciones, pero como veremos no hay problema alguno en modelar más niveles utilizando el mismo enfoque. En el nivel inferior aparecen los *conceptos*, que representan las unidades mínimas en las que se puede descomponer el conocimiento. El nivel inmediatamente anterior contiene los *temas*, que son agrupaciones de conceptos. Por último, aparecen las *asignaturas*, que son agrupaciones de temas. Consideraremos que todos los nodos son binarios, pero la interpretación que se da a la probabilidad de los distintos tipos de nodo es diferente: en los nodos concepto, representan la probabilidad de que el concepto se conozca o no se conozca, mientras que en los nodos asignatura y tema dicha probabilidad se interpreta como una medida del grado de conocimiento alcanzado en el tema y la asignatura. La justificación teórica que permite considerar las probabilidades de la forma descrita aparece en [Millán, Pérez-de-la-Cruz et al., 00].

Variables para recolectar evidencia. En nuestro caso serán preguntas tipo test multirespuesta. Las respuestas a dichas preguntas pueden ser correctas o incorrectas.

3.1.2 Enlaces

En esta sección vamos a definir las relaciones que se establecen entre las variables definidas. Respecto a las relaciones entre variables para medir el conocimiento, consideraremos que dominar un nodo de conocimiento tiene influencia causal en dominar aquellos nodos de conocimiento del nivel inmediatamente anterior en la jerarquía de granularidad que estén con él relacionados. En cuanto a la relación entre los nodos de conocimiento y las preguntas, consideraremos que poseer el conocimiento tiene influencia causal en responder adecuadamente a las preguntas. La red bayesiana resultante se muestra en la Figura 3, donde los nodos se han etiquetado de la siguiente forma: *A* representa al nodo asignatura, cada T_i representa un nodo tema, cada C_i representa un nodo concepto y cada P_i un nodo pregunta tipo test.

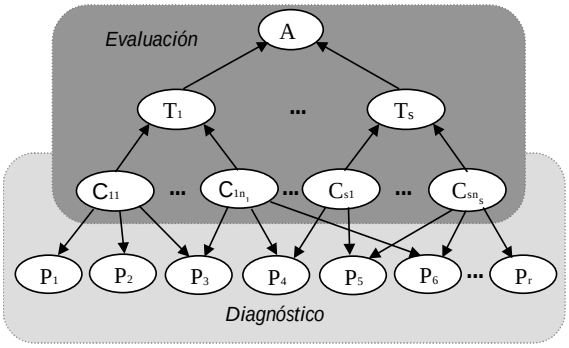

Figura 3. Red bayesiana para el diagnóstico mediante test

Como podemos apreciar, la red bayesiana se divide en dos partes que se solapan en los conceptos: la parte que soporta el proceso de diagnóstico, en el que se determina a partir de las respuestas del alumno el conjunto de conceptos que conoce/no conoce, y la parte que soporta el proceso de evaluación, en el que a partir de los resultados obtenidos en el proceso anterior se determina una medida del grado de conocimiento alcanzado por el alumno, tanto en la asignatura como en cada uno de los temas de los que consta. Cada una de estas partes se modela con un tipo de red bayesiana diferente: la parte de evaluación se modela con una red bayesiana clásica, mientras que para la parte de diagnóstico se utiliza una red bayesiana dinámica, puesto que en este caso es claro que los nodos tipo evidencia cambian con el tiempo, es decir, el hecho de que un alumno conteste correctamente a una pregunta relacionada con ciertos conceptos no quiere decir que siempre que le planteásemos una pregunta relacionada con tales conceptos la contestase también correctamente.

3.2 Parámetros

Definidas las relaciones, los parámetros que necesitamos especificar son:

Probabilidades a priori de los nodos concepto. Para ello podemos utilizar la información que haya disponible sobre el alumno en cuestión. En ausencia de información utilizaremos la distribución uniforme, es decir, consideraremos igualmente probable que domine el concepto o que no lo domine.

Probabilidades condicionadas de los temas dados los conceptos, y de la asignatura dados los conceptos. Estas probabilidades se pueden obtener de un modo sencillo a partir de un conjunto de parámetros más reducido: el conjunto de pesos que mide la importancia de cada concepto en el tema, y de cada tema en la asignatura. Para ello supongamos que C_{ij}, $i = 1, ..., n_j$, es el conjunto de conceptos relacionados con el tema T_j, y w_{ij} representa la importancia del concepto C_{ij} en el tema T_j, $i = 1, ..., n_j$. Entonces, las distribución de probabilidad conjunta se calcula de la siguiente forma:

$$P(T_j/(\{C_{ij}=1\}_{i\in S}, \{C_{ij}=0\}_{i\notin S}) = \frac{\sum_{i\in S}\omega_{ij}}{\sum_{i=1}^{n}\omega_{ij}}$$

para cada $S \subseteq \{1, ..., n_j\}$. De igual modo, para cada asignatura A sea $\{T_j / 1 \le j \le s\}$ el conjunto de temas relacionados, y para cada $j = 1, ..., s$ sea α_j el peso que mide la importancia relativa del tema T_j en la asignatura A. Entonces, la distribución de probabilidad condicionada de A se calcula mediante:

$$P(A/(\{T_i=1\}_{i\in S}, \{T_i=0\}_{i\notin S}) = \frac{\sum_{i\in S}\alpha_i}{\sum_{i=1}^{s}\alpha_i}$$

para cada $S \subseteq \{1, ..., s\}$.

Probabilidades condicionadas de cada pregunta dados los conceptos que en ella intervienen. Para ello, vamos a utilizar cuatro parámetros: c, que es el factor de adivinanza y representa la probabilidad de adivinar la respuesta correcta (es decir, $1/n$, donde n es el número de posibles respuestas); a, que es el nivel de dificultad de la pregunta; b, que es el índice de discriminación; y s que representa el llamado factor de descuido, es decir, la probabilidad de fallar la pregunta aún conociendo todos los conceptos que en ella intervienen. A partir de a, b y c definimos una función G mediante:

$$G(x) = \frac{(1-c)(1+e^{-1.7ab})}{1+e^{1.7a(x-b)}}$$

Esta función será utilizada para asignar las probabilidades de responder a la pregunta, de forma que cuantos más conceptos se sepan más probable será responder correctamente, como se ilustra en la Figura 4. Como podemos observar, las probabilidades se asignan de forma creciente, de forma que a mayor conocimiento más probable es contestar correctamente la pregunta. La implementación que hemos hecho permite también ordenar los conceptos por orden de importancia, de forma que la probabilidades más pequeñas se asignan a conocer los conceptos menos importantes.

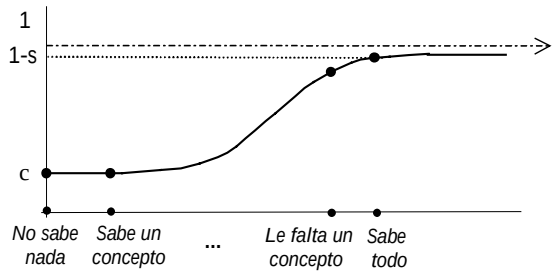

Figura 4. Probabilidades de contestar correctamente

3.3 Resultados

Para evaluar el funcionamiento del modelo propuesto, hemos utilizado alumnos simulados. El uso de alumnos simulados presenta las siguientes ventajas sobre el uso de alumnos reales: (a) es posible controlar totalmente las condiciones de la prueba de evaluación; (b) permite comparar los resultados obtenidos con los resultados reales, dado que es imposible disponer del verdadero estado cognitivo de un alumno real y (c) permite evaluar las técnicas antes de ser utilizadas con personas.

El funcionamiento de un alumno simulado es el siguiente: sean $C_1, ..., C_n$ los conceptos de la red de diagnóstico asociada a la asignatura que se pretende evaluar. Dado un valor $s \in [0,1]$, se define el *alumno simulado tipo s* como un alumno que conoce el $100s\%$ de los conceptos $C_1, ..., C_n$, donde el conjunto de los conceptos conocidos se genera aleatoriamente. De esta forma se obtienen alumnos simulados del mismo nivel pero cuyo conjunto de conceptos conocidos es diferente. Una vez generado el alumno simulado, se utiliza la red para calcular las probabilidades de responder correctamente a cada una de las preguntas. Dicha probabilidad se utilizará para simular el comportamiento del alumno en el test de la siguiente forma: supongamos que la probabilidad de responder correctamente a cierta pregunta P es p. Si el test plantea la pregunta P, se genera un número aleatorio a en el intervalo [0,1]. Si $p \ge a$, se considera que el alumno ha respondido correctamente a la pregunta, y si $p < a$, que la ha

respondido incorrectamente. Tras obtener la respuesta, el algoritmo de diagnóstico la utiliza para actualizar las probabilidades de los conceptos y elige la siguiente pregunta para proponerle al alumno. Como se ve este sencillo mecanismo permitirá comparar el diagnóstico obtenido tras la aplicación del test con el estado real de conocimiento del alumno.

Para las simulaciones hemos utilizado una red de pruebas compuesta por una asignatura *A*, cuatro temas T_1, ..., T_4, catorce conceptos C_1, ..., C_{14} y cien preguntas P_1, ... , P_{100}. Cada concepto se relaciona con uno o dos temas, según se representa en la Figura 5. A su vez, cada pregunta se relaciona con uno, dos o tres conceptos. Cada una de las preguntas las preguntas tiene seis respuestas posibles, y por tanto un factor de adivinanza de 1/6. Para cada nivel de dificultad el número de preguntas es aproximadamente el mismo, los factores de descuido son 0.001, 0.01 y 0.2 y los índices de discriminación son 0.2, 1.2 y 2 (el número de preguntas con cada factor de descuido y con cada índice de discriminación es aproximadamente el mismo).

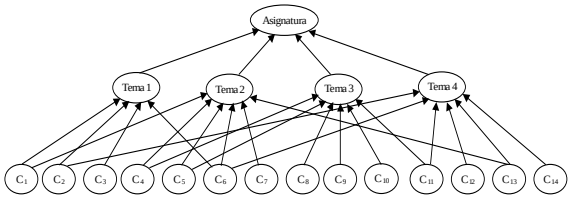

Figura 5. Relaciones entre conceptos, temas y asignatura

Se generaron 30 alumnos de seis tipos distintos: alumnos tipo 0.0, 0.2, 0.4, 0.6, 0.8 y 1.0. Esto hace un total de 180 alumnos. Para estudiar el comportamiento del algoritmo de diagnóstico, hemos procedido de la siguiente forma. En primer lugar fijamos un umbral *s* (en estas pruebas, 0.3). Tras finalizar el test, si la probabilidad de un concepto es mayor que 1-*s* consideramos que el alumno conoce el concepto, si es inferior a *s* que el alumno no lo conoce, y si queda entre ambos valores, que el algoritmo no ha sido capaz de diagnosticar si el alumno conoce o no el concepto. A cada alumno simulado le hemos realizado un test de sesenta preguntas, escogidas aleatoriamente entre las 100 preguntas disponibles. En la Figura 6 mostramos el número de conceptos (de un total de 14·180 = 2520 conceptos) que se diagnostican correcta/incorrectamente y que se dejan sin evaluar.

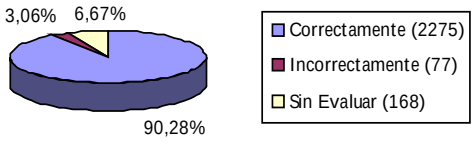

Figura 6. Resultados al final del test

Como vemos, el algoritmo diagnostica bien el 90.28% de los conceptos, mal el 3.06% y deja solamente el 6.67% de los conceptos sin evaluar. Teniendo en cuenta que el test consta de solamente sesenta preguntas, y que los conceptos diagnosticados son catorce, podemos calificar los resultados obtenidos como muy buenos. Sin duda ello se debe a la consistencia teórica del modelo utilizado.

En cuanto a la eficacia del proceso evaluador, es evidente que dependerá en gran medida de la eficacia del proceso de diagnóstico. El mecanismo del proceso evaluador es el siguiente: una vez finalizado el test, todos aquellos conceptos que han sido diagnosticados como sabidos se instancian a conocidos, y los que han sido diagnosticados como no sabidos a no conocidos. El resto de los conceptos conservan la probabilidad obtenida al final del test. Esta información se propaga en la parte correspondiente de la red, y de esta forma se obtienen las probabilidades de cada T_i, $1 \le i \le s$, y de *A*. Esta probabilidad puede considerarse como una estimación del grado de conocimiento alcanzado por el alumno en cada parte, y será comparada con el grado de conocimiento real que posee el alumno, que se calcula como la probabilidad obtenida al instanciar los conceptos que sabe a conocidos y los que no sabe a conocidos. El error cometido será la diferencia de ambos valores (estas probabilidades pueden ser convertidas en calificaciones tradicionales sin más que multiplicarlas por 100). La media de los errores por tema varía entre un mínimo de 0.0264 y un máximo de 0.0530, con desviaciones típicas muy pequeñas. En una escala decimal, el error variaría entre dos y cinco décimas, lo cual parece aceptable dado que el modelo admite que alumnos sin conocimiento den la respuesta correcta y alumnos con todos los conocimientos necesarios para responder a una pregunta la fallen.

4 Conclusiones y trabajo futuro

En las páginas precedentes hemos expuesto una visión panorámica de los trabajos teóricos y prácticos realizados por nuestro grupo con la finalidad de proporcionar herramientas de modelado

robustas y bien fundamentadas en el campo de de los STI. Creemos haber mostrado que tanto la teoría clásica de la TRI como las más recientes redes bayesianas son un punto de partida adecuado para ello.

Sin embargo, aún queda un largo camino por recorrer. Por ejemplo, será necesario integrar más estrechamente ambos enfoques para obtener una herramienta más versátil. Por otra parte, la evidencia que se tiene en cuenta en las versiones actuales de nuestros sistemas es muy reducida: únicamente las respuestas del alumno a las preguntas multirrespuesta. Para integrar realmente estas herramientas en un STI, será necesario tomar además en consideración otro tipo de datos, como pueden ser peticiones de ayuda del alumno y, en general, cualquier tipo de *episodios instructivos*. En esta línea se desarrollan nuestras actuales investigaciones.

Referencias

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

Conejo, R., Millán, E., Pérez-de-la-Cruz, J. L., & Trella, M. (2000). An Empirical Approach to On-Line Learning in SIETTE. In *Proceedings of 3rd International Conference on Intelligent Tutoring Systems ITS´2000*. LNCS 1839, Springer Verlag.

Flaugher, R. (1990). Item Pools. In H. Wainer (ed.), *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Hambleton, R. K. (1989). Principles and selected applications of Item Response Theory. In R. L. Linn (ed.), *Educational Measurement*. New York: MacMillan.

Huang, S. X. (1996). On Content-Balanced Adaptive Testing. In *Proceedings of 3rd International Conference CALISCE'96*. LNCS 1110, Springer Verlag.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (ed), *Computer assisted instruction, testing and guidance* (139-183). New York: Harper and Row.

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

Lord, F. M. &. N. M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Millán, E., Pérez-de-la-Cruz, J. L., & Suárez, E. (2000). An Adaptive Bayesian Network for Multilevel Student Modelling. In *Proceedings of 3rd International Conference on Intelligent Tutoring Systems ITS´2000*. LNCS 1839, Springer Verlag.

Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351-371.

Pearl, J. (1988). *Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: Danish Institute for Educational Research.

Ríos, A., Millán, E., Trella, M., Pérez-de-la-Cruz, J. L., & Conejo, R. (1999). Internet Based Evaluation System. In *Proceedings of the 9th World Conference of Artificial Intelligence and Education AIED'99*. IOS Press.

Shute, V. J. (1995). Intelligent Tutoring Systems: Past, Present and Future. In D. Jonassen (ed), *Handbook of Research on Educational Communications and Technology*. Scholastic Publications.

Thissen, D., & Mislevy, R. (1990). Testing Algorithms. In H. Wainer (ed.), *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

VanLehn, K. (1988). Student Modelling. In M. C. Polson, & J. J. Richardson (eds.), *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

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

Wainer, H., & Messick, S. (1983). *Principles of modern psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 12,* (361-375).