ESTUDIO COMPARATIVO ENTRE ALGORITMOS QUE MIDEN LA PRECISIÓN DEL SITEMA DE SELECCIÓN DE ÍTEMS PARA TEST ADAPTATIVOS COMPUTARIZADOS

VANESSA ESTHER PACHECO DE LA ROSA

TUTOR: LUZ STELLA ROBLES PEDROZO

UNIVERSIDAD TECNOLÓGICA DE BOLÍVAR

FACULTAD DE INGENIERÍA

PROGRAMA DE INGENIERÍA DE SISTEMAS

Cartagena de Indias, D.T. y C., 2017

Agradecimientos

A Jehová por su ayuda todos estos años;

A mi familia por sus buenos deseos;

A mi tutora por ser como una madre;

A mis profesores por su voto de confianza;

A mis amigos por su entusiasmo;

Índice de Contenido

Lista de Tablas

Lista de Figuras

Lista de Abreviaturas

CBR	Razonamiento Basado en Casos
EAI	Entornos de Aprendizaje Interactivo
EMV	Estimación por Máxima Verosimilitud
STI	Sistemas Tutoriales Inteligentes
TAI	Test Adaptativo Computarizado
TRI	Teoría de Respuesta al Ítem

1. Descripción del Proyecto

1.1. Introducción

Esta investigación busca implementar dos algoritmos de selección de preguntas en TAIs, a saber, el algoritmo basado en los conceptos débiles del alumno y el algoritmo basado en la función de información de un ítem, adaptados a los modelos logísticos de 1, 2 y 3 parámetros conocidos como modelo de *Rasch*, modelo *ltm* y modelo *tpm*, respectivamente. Posterior a su implementación, se han de comparar las preguntas seleccionadas por cada algoritmo y las debilidades del evaluado para determinar cuál algoritmo se ajusta mejor a sus deficiencias.

1.2. Objetivos

Objetivo general

Diseñar un estudio comparativo que mida la precisión entre el procedimiento de selección de preguntas a partir de las debilidades conceptuales del evaluado, y el método de selección de preguntas a través de la Función de Información, en Test Adaptativos Computarizados.

Objetivos específicos

- Identificar debilidades conceptuales en estudiantes evaluados a través de un test de estructura fija, con el fin de contar con un conjunto de registros de base sobre los cuales realizar las futuras pruebas de selección de ítems adaptativos.
- Diseñar un banco de ítems simulados, desde el cual se puedan realizar pruebas de selección de ítems adaptados al nivel de habilidad del sujeto, o ítems que evalúen conceptos coincidentes con las debilidades conceptuales del sujeto evaluado.
- Identificar y aplicar el modelo matemático que permite calcular la Función de Información para cada ítem en el banco de preguntas.
- Implementar el algoritmo que selecciona -teniendo en cuenta el perfil del usuario-, preguntas a partir de los conceptos débiles -identificados en el perfil-, y preguntas a partir de la Función de Información.

- Estimar el nivel de precisión, en porcentaje de respuestas correctas, obtenido al aplicar con ítems aleatorios, TAI con ítems a través de la Función de Información, TAI por debilidades conceptuales.

1.3. Resumen de los capítulos

A continuación, se muestra la estructura de este trabajo:

En el capítulo 2 se contextualiza al lector sobre la motivación que dirigió este proyecto y la problemática actual. Asimismo, se identificaron los trabajos que ya han hecho otros autores entorno a la temática a desarrollar y se presentaron conceptos importantes para el correcto entendimiento del proyecto.

En el capítulo 3 se describe menudamente la metodología empleada para llevar a cabo los objetivos de esta investigación. Se mencionan las herramientas a utilizar, los datos necesarios y los instrumentos de análisis que permitirán establecer una comparación entre los algoritmos estudiados.

Luego, en el capítulo 4 se explica detalladamente la implementación de los dos algoritmos seleccionados para estudio, incluyendo las bases de datos a construir.

En el capítulo 5 se detallan los procesos de pruebas y análisis de resultados. En las pruebas se aplican los algoritmos y se mide la precisión en la selección de ítems.

Por último, en el capítulo 6 se presentan las conclusiones y se proponen trabajos futuros.

2. Contexto y Motivación

2.1. Introducción

En este capítulo se explica detalladamente la pregunta que busca responder esta investigación y por qué esta fue seleccionada al describir el problema subyacente. Además, se puntualizan los algoritmos que en la actualidad se usan para la selección de ítems en TAIs y los trabajos investigativos que han abordado esta temática y han aportado soluciones a la problemática tratada en este proyecto.

2.2. Contexto y Descripción del problema

La evaluación del alumnado, específicamente, la evaluación final, pretende medir lo que los alumnos conocen, comprenden y son capaces de hacer, es decir, medir su nivel de aprovechamiento (Eurydice, 2006). Usualmente, los resultados de una evaluación final se utilizan para determinar si un alumno ha alcanzado un nivel suficiente para pasar a la próxima etapa en el proceso de aprendizaje. Siendo la evaluación un factor decisivo en el progreso de un estudiante, se ha de procurar que los tests realizados para confirmar que se domina lo aprendido sean lo más precisos posible, es decir, valoren con exactitud el desempeño del individuo.

Con este objetivo, se han diseñado tests adaptativos computarizados (CAT, también llamados TAI - Test Adaptativos Informatizados) con distintos algoritmos de selección de preguntas. Uno de los paradigmas de diseño de TAIs más utilizados es la Teoría de Respuesta al Ítem (TRI), y según este, las preguntas han de ser seleccionadas basándose en su Función de Información, esto es, la cantidad de información que el ítem aporta a la medida de habilidad. Sin embargo, un estudio realizado por Robles y Rodríguez (2012) propone diseñar TAIs cuya estimación de la habilidad del evaluado se haga en base a las debilidades conceptuales del individuo.

Por lo tanto, este proyecto pretende presentar un análisis comparativo entre estos procedimientos de selección de ítems con el propósito de responder la siguiente pregunta de investigación: ¿qué procedimiento de selección de preguntas se adapta mejor al desempeño del evaluado: selección de ítems a partir de las debilidades conceptuales del individuo o a través de la Función de Información?

2.3. Justificación

Según la definición de la Real Academia de la Lengua Española, un test es una prueba destinada a evaluar conocimientos o aptitudes, en la cual hay que elegir la respuesta correcta entre varias opciones previamente fijadas. Los test pueden ser orales o escritos, de estructura fija o flexible, dependiendo de qué se quiere evaluar, cómo se quiere evaluar y cuál es el contexto en el que se esté aplicando la evaluación.

Las evaluaciones son generalmente la forma más común y efectiva de evaluar el conocimiento o la habilidad de un alumno. Las evaluaciones tradicionales no siempre satisfacen la necesidad de discriminar el conocimiento de los alumnos, y los atributos tales como: *el tiempo de finalización del examen*, así como *el grado de dificultad de un test*, son difíciles de controlar.

Los test realizados a través de un computador -TAI, han demostrado ser más eficaces y eficientes que las pruebas tradicionales de papel y lápiz debido a varias razones:

- Usualmente los estudiantes deben responder un número reducido de preguntas al ser evaluados con TAIs, puesto que sólo se les administra ítems apropiados para su nivel de conocimiento, mientras que al responder un test tradicional un estudiante puede enfrentarse a preguntas muy difíciles o pasar rápidamente unas muy fáciles.
- Debido a que los TAIs se realizan en línea, los estudiantes tienen acceso a la retroalimentación del test en menos tiempo.
- Los TAIs proveen puntajes uniformemente precisos para la mayoría de evaluados (Thissen y Mislevy, 2000). En contraste, los tests tradicionales suelen ser muy precisos al evaluar estudiantes de habilidad media y poco exactos con estudiantes cuya habilidad es muy alta o muy baja.

Con la interactividad y la adaptabilidad del usuario, la aplicación de los TAIs amplía las posibilidades de realizar pruebas más allá de las limitaciones de las pruebas tradicionales de papel y lápiz.

Sin embargo, la aplicación de los TAIs tiene limitaciones. Una de las desventajas que presentan los Tests Adaptativos Informatizados es la selección de los ítems para un grupo de personas con la misma habilidad, mientras que otros ítems son rara vez escogidos; esto no sólo resultaría en la disminución de la dificultad del test, también la estimación de la habilidad del alumno estaría sesgada, la validez del test sería más baja y se afectaría la seguridad del banco de preguntas. Si un alumno fuese evaluado con un test que presente este inconveniente, el conocimiento que el evaluado realmente ha adquirido no podría ser medido con exactitud.

Para garantizar que al alumno evaluado se le están presentando los ítems directamente relacionados con sus debilidades conceptuales, evitar o disminuir el sesgo en la información que visualiza, y así mismo garantizar que los ítems presentados están directamente relacionados con sus falencia conceptuales, en esta investigación se propone: identificar los conceptos que un alumno no domina a partir de una respuesta incorrecta, diseñar y mantener un banco de preguntas de selección múltiple con única respuesta (con ítems que evalúen fuerte o muy fuertemente cada concepto), diseñar el algoritmo de selección de ítems basados en conceptos débiles, y establecer un comparativo que mida la efectividad de aplicar distintos métodos de selección de ítems (a partir del banco de preguntas). Este algoritmo de selección permite valorar lo que el alumno conoce y proporcionar preguntas guiadoras sobre los conceptos que no maneja.

Otro de los beneficios de utilizar este método de selección de preguntas es la retroalimentación. La retroalimentación es un elemento crucial del proceso de aprendizaje de un alumno y va de la mano con las evaluaciones. Uno de los principios para la buena evaluación y retroalimentación que Nicol (2007) presenta es: Aportar retroalimentación con información de alta calidad que ayude a los estudiantes a aprender de sus aciertos y errores. Al determinar los conceptos en los que un estudiante está fallando, este tiene la capacidad de enfocar su aprendizaje.

2.4. Marco Teórico

2.4.1. La Evaluación en la Educación

En este aparte se hará una descripción de los posibles entornos donde puede integrarse la metodología de retroalimentación que esta investigación toma como base.

2.4.1.1. Sistemas Tutoriales Inteligentes y su Evolución

En los STI, el alumno se enfrenta al proceso de aprendizaje mediante una guía de acciones previamente precisadas que corresponden a un modelo de tutorización individual o de 'uno a uno'. En la estructura de estos sistemas juegan un papel relevante dos elementos, que son:

- El modelo del estudiante, que se utiliza para representar lo que el sistema supone que el estudiante ha aprendido.
- El modelo pedagógico, que contiene el conocimiento referido a la forma de se gestiona el propio proceso de aprendizaje.

<u>Componentes básicos</u>

La arquitectura de un STI, cuenta con los siguientes componentes:

- Modelo del dominio. Corresponde a la respuesta sobre la que se enseña, y contiene el conocimiento sobre la materia que debe ser aprendida.
- Modelo del alumno. Proporciona información acerca del estudiante como individuo, el cual hace uso de un sistema de aprendizaje basado en computador. El sistema usa el modelo del estudiante para ayudar a determinar las acciones apropiadas para aquel estudiante.
- Modelo de instrucción. Responde a la pregunta 'Cómo se enseña'. Por representar estrategias de enseñanza, hace referencia a cómo el sistema debe mostrar el material educativo al alumno.
- Interfaz. A través de ella se lleva a cabo la interacción hombre-máquina.

Evolución

Con la intención de superar los problemas detectados en los STI, surgieron los Entornos de Aprendizaje Interactivos -EAI- y los Micromundos. En estos sistemas predomina el uso de imágenes de video y otras representaciones gráficas. Sin embargo, unas de las mayores limitaciones de los EAI están relacionadas con *problemas de evaluación* -como medida de los resultados del aprendizaje-. Esta limitación se encuentra en la dificultad para definir objetivos de aprendizaje, analizarlos con instrumentos, y aplicarlos para medir la eficacia en promover dichos objetivos. [Boticario y Guadioso, 2003].

Más adelante se dan otros desarrollos como son los Sistemas de Hipermedia, que se basan en presentar al usuario los contenidos en forma de material hipertexto. El usuario del sistema accede a la información de forma interactiva, y navega a través de la información. [Boticario y Guadioso, 2001] provee mayor información.

Sin embargo, estos sistemas presentan varios problemas relacionados con la utilización de hipertextos [Hammond, 1993]: desorientación del usuario en la navegación, y el grado de libertad en el uso de la información por parte de un alumno interfiere en el uso que otros alumnos hagan del sistema. Por lo tanto, se fueron introduciendo en dichos sistemas capacidades de adaptación como una manera de guiar al alumno en su navegación por los contenidos de que disponían.

Posteriormente, se produce una división progresiva del campo en diversas áreas: aprendizaje colaborativo, modelado del usuario, interfaces adaptativas, aprendizaje a través de la web, estandarización de contenidos y cursos, etc. Se hizo evidente que, para poder integrar los diversos trabajos realizados en cada uno de estos campos, era necesario introducir mayor metodología y concretar formatos y estándares de control en el desarrollo de este tipo de aplicaciones.

2.4.1.2. El Diagnóstico en los Sistemas Tutoriales Inteligentes

Los sistemas que a continuación se describen hacen parte de un grupo de sistemas que se dividen en subgrupos en función de cómo llevan a cabo el diagnóstico.

Modelos de Evaluación basados en Distancia Semántica

El trabajo realizado por [Huapaya et al., 2007] propone una metodología de evaluación basado en la distancia semántica -DistSem-. DistSem se basa en las distancias estimadas entre significados, y refleja esta información en una matriz de similitudes. En la evaluación basada en DistSem, las preguntas de evaluación consisten en la asociación de conceptos propios del tema que se evalúa, es decir, cada estudiante da un nivel de asociación a un par de conceptos, por ejemplo: 1 – poco relacionados y 7 – muy relacionados. De acuerdo a las respuestas dadas por el estudiante, se genera una matriz de similitudes que refleja la red semántica del estudiante y por otra parte se construye la red semántica del experto para realizar la comparación y así poder diagnosticar el nivel de conocimiento del estudiante [Huapaya et al., 2007].

Modelos de Evaluación basados en Redes Neuronales

Tal como lo describe [Roa et al., 2005], esta presenta un acertado procedimiento para modelar al estudiante y así poder adaptar los contenidos de un curso virtual -en la plataforma Moodle-, a las características del estudiante. El proceso de adaptación, del curso al perfil del estudiante, se realiza de acuerdo con las acciones que el usuario va ejecutando durante las actividades propias del curso y la interacción monitoreada que el sistema hace sobre el estudiante.

Modelos de Evaluación basados en Mapas Conceptuales

Los mapas conceptuales son una estrategia que permite hacer explícito el conocimiento conceptual que un estudiante ha obtenido al estudiar un dominio en particular. La técnica de evaluación adaptativa basada en mapas conceptuales reemplaza los test por mapas conceptuales, y tienen como fin diagnosticar habilidades de orden superior [Anohina et al., 2007].

En este sistema, lo que hace adaptable la evaluación es el hecho de presentarle al -estudiante-, mapas conceptuales de acuerdo al nivel de dificultad que él puede afrontar, así como presentarle ayudas cuando el sistema detecta que el estudiante lo necesita.

Modelos de Evaluación basados en Test Adaptativos

A continuación, se describe un modelo de evaluación que está entre los más conocidos y que hace uso de test computarizados para generar una estimación del nivel de habilidad de un sujeto evaluado:

HEZINET

Este modelo es un sistema educativo adaptativo para la web, orientado a la enseñanza de la lengua vasca, desarrollado e implementado en diversos centros de educación [Gutiérrez; López- Cuadrado et al., 2002]. Este sistema tiene un banco con alrededor de 600 ítems. Trabajan con la Teoría de Respuesta al Ítem, con el modelo logístico de tres parámetros, y respuesta dicotómica. En cuanto a la calibración de los ítems, este sistema lo lleva a cabo de manera convencional -los aplica previamente y determina cuáles son adecuados-, y una vez seleccionados hace uso de ellos en los test adaptativos. Sin embargo, no está muy claro cómo se hace la selección de ítems que evalúen todo el contenido del curso o del tema objeto de evaluación.

Modelos de Evaluación basados en Redes Bayesianas

Los sistemas de evaluación y tutorización basados en redes bayesianas no abundan, y los pocos que las implementan permiten representar distribuciones de probabilidad conjuntas que requieren menos espacio para ser almacenadas comparados con una representación tabular [VanLehn y Martín, 1998]. Algunos son: OLAE Y SIETTE, HYDRIVE y ANDES.

2.4.1.3. Los Test Adaptativos Informatizados

La definición del concepto de Test Adaptativo más usada es la aportada por [Wainer, 1990]: '*La idea fundamental de un test adaptativo es imitar de forma automática el comportamiento de un examinador. Esto es, si un examinador le presenta al alumno un ítem demasiado difícil para él, éste dará una respuesta errónea, y por lo tanto, la siguiente vez, el examinador presentará una pregunta algo más fácil*'.

Generalmente, un Test Adaptativo comenzará presentando una pregunta de nivel medio. Si el examinando la acierta, la siguiente será algo más difícil. Si por el contrario falla, la siguiente será más fácil [Guzmán, 2005]. Un Test Adaptativo Informatizado -TAI- [Wainer, 1990; Olea et al., 1999] es una herramienta de medida administrada a los alumno por medio de un computador.

En términos más precisos, un TAI es un algoritmo iterativo que comienza con una estimación inicial del conocimiento del alumno y continúa con los siguientes pasos:

- Todos los ítems, que no han sido administrados todavía, son analizados para determinar cuál de ellos contribuye en mayor medida a una mejor estimación del conocimiento del alumno.
- El ítem se muestra al alumno.
- En función de la respuesta elegida por el examinando, se estima el nuevo nivel de conocimiento de éste.
- Los pasos del 1 al 3 se repiten hasta que el criterio de finalización del test se satisfaga.

2.4.2. La Evaluación mediante Test

A continuación se hace una descripción de las fases que son recomendables seguir en la construcción de un test de evaluación.

Primera Fase: Redacción de Preguntas

Para este caso, en el que el docente decide redactar las preguntas del cuestionario que va a aplicar a sus estudiantes, se le recomienda en términos generales, lo siguiente:

- Evitar el uso de palabras tales como: siempre, generalmente, no, ninguno.
- Evitar negaciones dobles.
- Evitar errores gramaticales y ortográficos.

- Expresar la idea principal de la pregunta en el enunciado.
- La dificultad de la pregunta no debe estar representada en la comprensión del enunciado.
- Hacer enunciados precisos, puntuales, evitando colocar información irrelevante.
- La redacción del enunciado y su pregunta deben conducir a solo una opción como respuesta, y no permitir divergencias.
- No dejar en el enunciado pistas que sugieran cuál es opción de respuesta correcta.
- Los enunciados de las preguntas no deben ser excesivamente largos.
- El vocabulario utilizado en los enunciados de las preguntas debe estar al nivel de los evaluados.
- Evitar elementos o gráficos que puedan confundir al evaluado.
- Evitar redacción que pueda herir susceptibilidades con enunciados con sesgo racial, étnico, o de género.

Segunda Fase: Construcción de un banco de ítems

Un banco de ítems es un conjunto de ítems que mide una misma variable. La construcción de un banco de ítems se hace con la finalidad de contar con preguntas que formen parte de futuros test, y adicionalmente, que puedan luego ser utilizados en tests administrados en soporte informático.

En general, los ítems que forman parte de un banco de ítems deberían tener presente:

- Claramente identificado el rasgo o característica que está midiendo.
- Desde el punto de vista jerárquico qué temas y conceptos evalúan.
- Especificada la población hacia la cual va dirigida.

– Determinación del sistema de anclaje a utilizar, que puede ser: Anclaje de ítems -los sujetos de todas las muestras responden a un conjunto de ítems-, Anclaje de sujetos -todos los ítems se aplican a un grupo de sujetos-.

Lo ideal es que cuando un ítem forme parte de un banco de ítems, ya se le hayan estimado sus parámetros, en cuanto a indicar qué tan difícil es, o cuál es la probabilidad de que un estudiante que no tenga claro cuál es la respuesta correcta entonces adivine, etc., pero para ésto, un test completo ya ha debido ser aplicado en campo, solo así esta información puede ser suministrada a cada ítem del banco de ítems, y amarrado a un test en específico.

Tercera Fase: Calibración del banco de ítems

Cuando se habla de estadísticas que describan a los ítems, se está haciendo referencia a valores que den cuenta de qué tan difícil es el ítem para los estudiantes -este valor se conoce ya al final de haber sido aplicada una prueba-, qué tantas opciones de respuestas distractoras o absurdas fueron seleccionadas en un ítem, qué tanto un ítem puede ayudar a determinar si un estudiante tiene claro o no un concepto, etc. El proceso de estimación de estos estadísticos o parámetros de los ítems, se llama calibración, y la base teórica que hace necesario el cálculo de los mismos -en este trabajo-, es la Teoría de Respuesta al Ítem -TRI-.

Cuarta Fase: Aplicación del Examen - Test

Una vez redactados los ítems o preguntas del examen, la aplicación de éste puede hacerse de la manera tradicional, o digital haciendo uso del computador. Interesa para el presente trabajo hacer una simulación de la aplicación de un TAI.

2.4.2.1. Fiabilidad y Validez de un Instrumento de Medida

El diseño de un test debe contener, al menos, las siguientes características [Yela 1996]:

– Fiabilidad, que hace referencia a: que las puntuaciones o resultados de un test, obtenidos en una determinada ocasión, y bajo ciertas condiciones, se mantengan estables en el tiempo.

- La validez, que hace referencia a: que realmente se esté midiendo lo que el test dice medir. En otras palabras, la validez está referida al grado en que cada prueba refleja el *constructo* que dice medir.

- Tipicidad o baremos, que hace referencia a: valores normativos de la población con los cuales comparar el puntaje o resultado de una persona -obtenido a través de una prueba-, miembro de esta población.

Análisis de fiabilidad del instrumento

Para estimar la *Confiabilidad de Consistencia Interna*, la cual consiste en determinar el grado de correlación que existe entre los ítems de una misma prueba, existen diferentes procedimientos, algunos de ellos son:

- Kuder-Richardson [Kuder-Richardson 1937]
- Alpha de Cronbach [Alpha de Cronbach 1951]
- Rulon [Rulón 1939]
- Hoyt [Hoyt 1941]

De los procedimientos para la estimación de la consistencia interna antes descritos, se recomienda el Alpha de Cronbach en [Robles 2012], porque los ítems del modelo de evaluación que contempla esta investigación, tienen varias opciones de respuesta, y de acuerdo con la teoría de factores que afectan la estimación de la fiabilidad, esto incide de manera positiva sobre la magnitud del coeficiente de fiabilidad que se va a estimar. En esta investigación se empleará el alfa de Cronbach para medir la confiabilidad.

Validez

Si bien todos los aspectos relacionados con garantizar que un examen o test es adecuado para ser aplicado, y adicionalmente, el que sea confiable, se constituyen como elemento mínimo para poder ser

utilizado correctamente [Yela, 1996], el requisito más importante de un test es su validez, debido a que si éste no mide lo que dice medir, no sirve de nada tener fiabilidad.

Para construir el test de estructura fija del cual se obtuvieron los resultados que conforman los datos d entrada de esta investigación se desea midieron la *Validez de Contenido* y la *Validez de Constructo*.

La *Validez de Contenido* se estima de manera subjetiva, de ahí que se requiera de juicios de expertos, conocedores del tema de dominio, que pueden dar cuenta de: la claridad de la redacción de los ítems, y la consistencia entre la intención plasmada en el ítem versus el dominio o contenido.

Con respecto a la *Validez de Constructo*, la validez de constructo interesa cuando se quiere utilizar el desempeño de los sujetos con el instrumento, para inferir la posesión de ciertos rasgos o cualidades psicológicas [Gronlund, 1976].

Para evaluar la Validez de Constructo de un test, existen varias aproximaciones, una de ellas es el Análisis de Factores. Un Análisis Factorial es el análisis que se realiza sobre la matriz de correlaciones entre los ítems, con el fin de descubrir estadísticamente los factores y sus elementos [Abad et al., 2006]. A la muestra de datos del estudio previo a esta investigación se le aplicó Análisis Factorial para comprobar si la estructura resultante -grupos de factores identificados-, coincidía con la estructura teórica -asumida a priori-, y de esta manera, confirmar el modelo teórico.

2.4.2.2. Base Teórica de la Teoría Clásica de Test versus la Teoría de Respuesta al Ítem

La Teoría Clásica de Test

Esta teoría tiene sus inicios a comienzos del siglo XX, con Spearman en [1904, 1907, 1913]. Spearman propone un modelo simple para las puntuaciones que obtienen las personas luego de aplicar un test, éste consiste en asumir que la puntuación que obtiene una persona en un test -llamada puntuación empírica-, está formada por dos componentes: una puntuación que obtiene la persona luego de aplicar a un test, y un error; este error puede deberse a muchas causas que en la mayoría de los casos no se puede controlar.

Debilidades de La Teoría Clásica de Test

Habían dos cuestiones básicas que no encontraban buena solución en la teoría clásica y que hacían que la medición en el campo de la psicología no fuera homologable a la medición que exhibían otras ciencias empíricas [Muñiz, 2010], son ellas:

- Tal como Muñiz lo explica en [Muñiz, 2010], si se evalúa la inteligencia de tres personas distintas con un test diferente para cada persona, los resultados no son comparables, no se puede decir qué persona es más inteligente porque no se tiene la misma base de medida.
- Como lo explica [Muñiz, 2010], propiedades psicométricas importantes de los test, tales como la dificultad de los ítems o la fiabilidad del tests, están en función del tipo de personas utilizadas para calcularlas.

Estos aspectos o dificultades son solventados por la TRI.

La Teoría de Respuesta a los Ítems

El objetivo principal de la TRI es conseguir medidas invariantes respecto de los sujetos medidos y de los instrumentos utilizados [Muñiz, 1997].

Supuesto de Unidimensionalidad

Para la TRI, el supuesto de unidimensionalidad exige que el rendimiento en un ítem dependa del nivel del evaluado en un solo rasgo. Sin embargo, el cumplimiento del supuesto de unidimensionalidad nunca se cumple totalmente porque el rendimiento en un test está influido por variables cognitivas y de personalidad tales como la motivación y en algunos casos la ansiedad, etc. Una medición bruta de la unidimensionalidad es el coeficiente alfa, porque éste mide la coherencia interna de los ítems en un test, sin embargo un método más adecuado para evaluar la unidimensionalidad de un test es el análisis factorial [Hambleton y Rovinelli, 1986]. Si el análisis factorial revela la existencia de un solo factor dominante, es éste un buen soporte para probar la Unidimensionalidad.

Supuesto de Independencia Local

Este supuesto de independencia local está referido a los ítems; y por tanto, se dice que existe independencia local entre los ítems, si la respuesta que un sujeto da a un ítem no depende de las respuestas que da a los otros.

Modelos de la TRI

En la TRI se han desarrollado un conjunto de modelos matemáticos en los que se asume que la probabilidad de que una persona emita una determinada respuesta ante un ítem puede ser descrita en función de la posición de la persona en el rasgo o aptitud latente (θ), y de una o más características de ítem (índice de dificultad, de discriminación, probabilidad de acertar por azar...).

Modelos Básicos

Actualmente, los modelos logísticos son los más populares y a su vez pueden ser clasificados en función del número de parámetros que los caracterizan, es así como existen los modelos logísticos de un parámetro -1PL-, de dos parámetros -2PL-, y de tres parámetros -3PL-. Estos modelos tienen en común el uso de una función matemática para especificar la relación entre el desempeño observable del evaluado en un test y los rasgos o habilidades no observables que se supone están implícitas en el desempeño en el test.

Modelo Logístico de un Parámetro

Este modelo es el más simple de todos, también llamado modelo Rasch [Rasch, 1960]. La probabilidad de acertar un ítem solamente depende del nivel de dificultad del ítem y del nivel del sujeto, -nivel de rasgo o habilidad-, en la variable que es medida. Los parámetros en el modelo Rasch se estiman con el método de máxima verosimilitud, que consiste en determinar los parámetros que hacen más probable las respuestas observadas. La función matemática que define el modelo es la siguiente:

$$P(\theta) = \frac{e^{D(\theta - b)}}{1 + e^{D(\theta - b)}} = \frac{1}{1 + e^{-D(\theta - b)}} \qquad \text{(III.1)}$$

Donde:

P(θ): Es la probabilidad de acertar el ítem, si el nivel de rasgo es θ

θ: Nivel de habilidad del sujeto

b: Índice de dificultad del ítem

e: Base de los logaritmos neperianos (2.718)

D: Constante (D = 1.7 o 1)

El nivel de habilidad (θ) de un sujeto puede definirse en cualquier escala, sin embargo, es normal utilizarse una escala con media cero, varianza uno, y un intervalo de valores entre -3.0 y 3.0. En este proyecto se usará una escala del 0 al 100.

Modelo Logístico de dos Parámetros

Este modelo añade al anterior un segundo parámetro 'a' que indica la capacidad discriminativa del ítem, o capacidad que tiene el ítem de distinguir entre un sujeto hábil y uno menos hábil. La función matemática que define el modelo es la siguiente:

$$P(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} = \frac{1}{1 + e^{-Da(\theta - b)}} \qquad \text{(III.2)}$$

Donde:

a: Índice de discriminación

Modelo Logístico de tres Parámetros

Este modelo adiciona al modelo anterior, un tercer parámetro 'c' que representa la probabilidad de acertar un ítem al azar. En otras palabras, representa la probabilidad de que sujetos de aptitud muy baja contesten correctamente a un ítem -adivinen la respuesta correcta-. La función matemática que define el modelo es la siguiente:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}} \qquad \text{(III.3)}$$

Donde:

c: Factor de adivinanza

Estimación de Parámetros – Calibración de Ítems

Una vez se ha seleccionado el modelo de la TRI, se debe continuar con la aplicación del test a grupo de sujetos, y a partir de la matriz de respuestas obtenidas, estimar los parámetros de los ítems -bien sea a, b, c-, dependiendo del modelo utilizado -1P, 2P o 3P-, y los parámetros del rasgo latente (θ) de los sujetos.

Los métodos para estimar parámetros de modelos de la TRI se basan fundamentalmente en el principio de máxima verosimilitud, en criterios bayesianos o en estrategias heurísticas. El método de máxima verosimilitud busca los valores de los parámetros que hacen más probable la obtención de los datos empíricos a partir del modelo. Este método será implementado en este proyecto para hacer el cálculo de la aptitud.

Estimación por Máxima Verosimilitud

Para obtener el nivel de rasgo θ se aplica el *método de máxima verosimilitud*. Se desea encontrar los valores de los parámetros que hagan más probable la matriz de respuestas obtenida luego de aplicar un test o responder un ítem. Esta probabilidad según la EMV se describe de la siguiente forma:

$$L = \prod P^R Q^{1-R} \tag{III.4}$$

Donde P es la probabilidad de acierto según el modelo aplicado; Q, la probabilidad de error $(1 - P)$; R, el resultado en cada ítem (1 es acierto, 0 es error).

Función de Información

La función de información de un test expresa la precisión de las estimaciones de los valores obtenidos en el rasgo latente, como una combinación de la precisión de cada ítem. Únicamente se requiere conocer los parámetros de los ítems que conforman el test, y el rango de valores en ese rasgo, que caracterice a la población objetivo para calcular la IIF.

La fórmula para calcular la función de información de un ítem según el modelo *tpm* es la siguiente:

$$I(\theta) = a^2 \frac{Q(\theta)}{P(\theta)} \left[\frac{P(\theta)}{1-c}\right]^2 \tag{III.5}$$

Donde θ es la habilidad del evaluado, a es el valor del discriminante del ítem, $P(\theta)$ es la probabilidad de acertar el ítem, cuya fórmula varía dependiendo del modelo logístico a aplicar, $Q(\theta)$ es $1 - P(\theta)$ y c es el factor de adivinanza del ítem.

La fórmula de la función de información del ítem para los modelos de 1 y 2 parámetros se reduce a:

$$I(\theta) = a^2 P(\theta) Q(\theta) \tag{III.6}$$

2.4.3. Razonamiento Basado en Casos

Sistemas de razonamientos basados en casos SRBC: Modelo de razonamiento que permite resolver problemas, entender situaciones y aprender utilizando mecanismos de memorización, problemas superpuestos y criterios de optimalidad.

Los SRBC se sustentan en tres principios básicos:

- Solución de problemas superpuestos: se aplica en casos que utiliza casos resueltos menores.
- Principio de optimalidad de Bellman: memoriza la mejor solución, luego de un proceso de selección.
- Memorización: memoriza las soluciones obtenidas en la librería de casos para uso posterior.

Los SRBC presentan algunas ventajas frente a los sistemas tradicionales:

- Adquisición de conocimiento: La adquisición del conocimiento se realiza a partir de la experiencia previa almacenada en la librería de casos. Esto evita el proceso de extracción de reglas a partir de un experto en dominio, lo cual ya constituía de por si en un "cuello de botella" y además no se garantizaba la validez de las reglas.
- Mantenimiento del conocimiento: Los SRBC permite incrementar nuevos casos a la librería de casos sin la intervención del experto, haciendo innecesario el proceso de mantenimiento de la base de conocimiento en SRBR (Sistemas de razonamiento basados en reglas) que resulta costoso.
- Eficiencia en la resolución de problemas: la reutilización es un principio básico de la informática. Los SRBC permiten que se puedan resolver casos similares sin tener que rehacer la base de conocimientos.
- Calidad de la solución: Al aplicar el principio de optimalidad, se garantiza memorizar la mejor solución o lo que ha sucedido en un contexto determinado.

La construcción de sistemas inteligentes simula de algún modo la manera en que los seres humanos resuelven problemas. Los Sistemas de razonamiento basado en casos (SRBC) parten de un problema ya

resuelto (caso) alojado en una librería de casos. Estas tareas son las que usualmente se realizan en la vida cotidiana. La reutilización de software es una de las aproximaciones más efectivas de los SRBC. Se trata de aprender de ejemplos, casos o datos conocidos de tal forma que se tome una decisión sobre nuevos casos. Desde el enfoque de minería de datos, en tareas de clasificación, podemos asignar una clase a un nuevo caso observando las clases similares. Igualmente, en tareas de agrupamiento, asignaremos un nuevo ejemplo al grupo donde estén los casos más similares. En el caso de regresión, el valor predicho para un nuevo ejemplo no puede distar mucho de los valores obtenidos para ejemplos similares. Las líneas de producto software generalmente surgen a partir de aplicaciones existentes. Una organización tiene una librería de aplicaciones (casos) y cuando se requiere una nueva aplicación, se utiliza la primera como base para la nueva aplicación. Se propone la utilización de Sistemas de razonamiento basados en casos para el desarrollo de líneas de software.

2.5. Estado del Arte

Son dos las teorías que guían la construcción de la mayoría de los tests y otros autores las han comparado para mostrar las ventajas y desventajas que tiene la una sobre la otra, confrontación que permite elegir qué teoría se ajusta mejor a las necesidades e intereses de un evaluador. Otro factor que afecta la precisión de un test adaptativo informatizado es el algoritmo usado para la selección de ítems. Es por esto que varios proyectos investigativos han sido orientados a la implementación de distintos algoritmos. Además, esfuerzos se han hecho para personalizar el aprendizaje al tener en cuenta el nivel de habilidad del evaluado y así ofrecer retroalimentación en las áreas necesarias.

Las Teorías de los Tests: Teoría Clásica y Teoría de Respuesta a los Ítems

Según Muñiz, J. (2010), el enfoque clásico posee limitaciones que se ven superadas dentro del marco de la Teoría de Respuesta al Ítem. Dos cuestiones básicas que no encuentran solución en este enfoque son: las mediciones no resultan invariantes respecto al instrumento utilizado, es decir, los resultados de diferentes tests que evalúan la misma habilidad no son comparables, pues cada test tiene su propia escala y, la ausencia de invarianza de las propiedades de los tests respecto de las personas utilizadas

para estimarlas; en otras palabras, el valor de los parámetros importantes en un test como la dificultad de los ítems está sesgado por el juicio de las personas utilizadas para calcularlos. Otra asunción de la Teoría Clásica, que la experiencia refuta, es que un test mide con la misma fiabilidad a todas las personas evaluadas. La TRI resuelve estos problemas graves del marco clásico al presentar una función llamada Curva Característica del Ítem (variable medida vs. probabilidad de acertar el ítem) determinada por el valor de tres parámetros: el índice de discriminación del ítem, la dificultad del ítem y la probabilidad que hay de acertar un ítem al azar, y al suponer que los ítems deben ser unidimensionales e independientes unos de otros.

A Personalized Assessment System based on Item Response Theory

El artículo de Lee, Y., Cho, J., Han, S., & Choi, B. (2010) se enfoca no solamente en estimar el nivel de habilidad de un estudiante, sino en identificar un número determinado de características de éste, así como en proveer retroalimentación en las áreas que el evaluado necesita mejorar. Con este objetivo, presenta un sistema de tutorías inteligente (ITS, por sus siglas en inglés) para aprender inglés disponible en la web. Debido a que ITS está basado en inteligencia artificial, este provee un entorno de aprendizaje dinámico y flexible que facilita el análisis de parámetros de los ítems y de las respuestas de cada evaluado, y resulta en un aprendizaje personalizado, pues la retroalimentación se basa en el nivel de conocimiento o habilidad de cada estudiante evaluado.

Self-Assessment in a Feasible, Adaptive, Web-Based Testing System

Guzmán, E. & Conejo R. (2005) presentan una solución web para generar y construir TAIs llamada SIETTE (Sistema de Evaluación Inteligente mediante Tests). La arquitectura de este sistema está conformada principalmente por cursos o materias, un repositorio de modelos de estudiantes donde se guarda la información de cada alumno, un generador de tests y una herramienta para la calibración de ítems. Un profesor puede elegir entre dos criterios para escoger las preguntas presentadas a los alumnos: el criterio de información bayesiano o el criterio basado en la dificultad del ítem comparado con el nivel de habilidad del estudiante.

Ability Assessment based on CAT in Adaptive Learning System

Con el objetivo de evaluar de manera más eficiente el nivel de habilidad de un aprendiz, Chen, S. & Zhang, J. (2008) proponen integrar un módulo de tests adaptativos computarizados con un sistema de aprendizaje adaptativo. Por lo tanto, la arquitectura presentada por los autores está compuesta por los elementos del módulo de TAI, a saber, un banco de preguntas, un módulo de generación de ítems, una interfaz para la evaluación adaptativa y una base de datos con los resultados de los tests, y por los componentes de un sistema de aprendizaje adaptativo, específicamente, el modelo del dominio, el modelo de estudiante, el modelo de instrucción, el modelo adaptativo y la interfaz de usuario. Respecto a la selección de ítems, se expone en el artículo un modelo logístico de tres parámetros de la Función de Información del Ítem (IIF, por sus siglas en inglés). Los tres parámetros del modelo son: el discriminante, el nivel de dificultad y el factor de adivinanza de un ítem. El valor de esta función es calculado para todos los ítems antes de elegir una pregunta para el evaluado y el ítem con la IIF de mayor valor será presentado al estudiante, y su habilidad será determinada con el método de estimación por máxima verosimilitud. Adicionalmente, el criterio de parada recomendado por los autores se basa en la precisión de medida del test; cuando el error estándar es igual o menor que 0.33, el test ha de terminar.

Modeling the Examinee Ability on the Computerized Adaptive Testing Using Adaptive Network-Based Fuzzy Inference System

Lin, Chen&Tsai (2008) plantean la hipótesis de que el Sistema Adaptativo de Inferencia Difusa basado en la Red (ANFIS, por sus siglas en inglés) puede ser utilizado para inferir el nivel de habilidad de un estudiante. Este estudio presenta un modelo basado en ANFIS que puede escoger preguntas de forma adaptativa teniendo en cuenta la Teoría de Respuesta al Ítem. Antes de que el evaluado responda una pregunta, se tienen en cuenta el discriminante, la dificultad y el factor de adivinanza del ítem, así como el nivel de habilidad estimado del estudiante, y se escoge el ítem con mayor función de información; los factores antes mencionados permiten que el modelo propuesto por los autores infiera la habilidad del alumno y mejore este estimado. Las simulaciones hechas para evaluar el rendimiento de este

acercamiento muestran que este método es más eficiente que la estimación por máxima verosimilitud y la estimación por verosimilitud Bayesiana cuando el valor de la información del test está dentro de cierto rango de precisión.

User ProfilingbasedAdaptive Test Generation and Assessment in E–LearningSystem
En este artículo, Jadhav, Rizwan&Nehete (2013) describen un sistema de evaluación del perfil del usuario basado en aprendizaje electrónico con generación y valoración de TAIs. Este sistema permite que un estudiante tome un test luego de haber estudiado cierto concepto. Si al finalizar el test el estudiante presenta deficiencias en el concepto evaluado, el sistema analizará los tipos de errores cometidos por el alumno y en base a éstos seleccionará un nivel apropiado para el estudiante (B, básico; I, intermedio; A, avanzado) y creará un test de n preguntas seleccionadas al azar de cada clase de error cometido dependiendo de la proporción de cada error. Las implementaciones de este sistema muestran que éste reduce el esfuerzo que un alumno ha de hacer para aprender un concepto. Además, este sistema asegura que el usuario en realidad domina los conceptos en los que una vez tuvo dificultad.

2.6. Conclusiones

La TRI ha estado en constante evolución, buscando métodos para mejorar la evaluación de los estudiantes. Entre los elementos que se han de optimizar está el algoritmo de selección de ítems. Por lo tanto, el estudio hecho en esta investigación proporcionará conocimiento útil para dar soluciones a esta problemática.

3. Metodología

3.1. Introducción

El propósito de este capítulo es describir a grandes rasgos los pasos a seguir para cumplir con los objetivos de esta investigación. Trazar claramente el camino por recorrer informa al investigador del conocimiento y las tecnologías necesarios a lo largo de los procesos de indagación, implementación y análisis de resultados.

3.2. Detalles del procedimiento propuesto

El objetivo principal de este proyecto investigativo es presentar un estudio comparativo entre dos métodos de selección de ítems en TAI, a saber, el método basado en las debilidades conceptuales del evaluado y el método basado en la Función de Información del ítem, para determinar cuál es más preciso. Por ende, se diseñarán dos algoritmos de selección de ítems para implementar los métodos mencionados sobre un banco de preguntas simulado. Luego de su implementación, se espera identificar el procedimiento que evalúe con mayor exactitud el desempeño de un alumno.

En la primera etapa de la investigación se aplicarán tests de estructura fija de 20 preguntas para determinar las debilidades conceptuales de un grupo simulado de evaluados y con los resultados se formará un conjunto de registros base. Este conjunto de datos permitirá calibrar el banco de preguntas y formar el perfil de usuario de cada evaluado necesario para implementar el algoritmo de selección de preguntas basado en los conceptos débiles del alumno.

En la segunda etapa se construirá un banco de preguntas simulado cuyos campos serán: identificador (número entero), descripción, nivel de dificultad, determinante, factor de adivinanza, conceptos evaluados (estos conceptos tendrán un peso de 4) y fuertemente evaluados (conceptos con peso 5). Habrá tantas preguntas como combinaciones posibles de conceptos con pesos 4 y 5. Estas combinaciones se calcularán teniendo en cuenta que las preguntas pueden evaluar máximo 2 conceptos. A partir del conjunto de registros base se calcularán los parámetros de los ítems, específicamente, el nivel de dificultad, la probabilidad de adivinanza y el discriminante, proceso

llamado calibración del banco de preguntas. En la simulación, el estudiante responderá de forma correcta o incorrecta aleatoriamente. Para la creación del banco de preguntas y la simulación se utilizarán los lenguajes de programación Python y R respectivamente.

Como tercer paso se investigará sobre la Función de Información y después de haber identificado claramente su modelo matemático, éste será aplicado sobre cada ítem en el banco de preguntas.

En la cuarta etapa de la investigación se implementarán los algoritmos de selección de preguntas basados en la Función de Información y en las debilidades conceptuales del evaluado. Se hará la simulación de evaluar a un estudiante aplicando un test adaptativo informatizado que valore su nivel de habilidad a través de la función de información. Luego, se hará lo mismo con un test que valore el nivel de habilidad del alumno basándose en sus conceptos débiles. Cada ítem presentado será tomado del banco de preguntas simulado, evaluando acierto y desacierto en cada respuesta como base para presentación de la siguiente pregunta, e implementando los mecanismos de parada que permiten generar el nivel de desempeño del evaluado. Se diseñarán ambos modelos utilizando el lenguaje de programación Python.

Por último, los resultados obtenidos, a saber, los porcentajes de respuestas correctas al implementar cada algoritmo, serán comparados para determinar la precisión de los modelos, siendo el algoritmo con mayor porcentaje el más preciso.

3.3. Conclusiones

Para cumplir con los objetivos definidos con anterioridad será necesario aplicar un test de estructura fija para determinar los conceptos débiles de un conjunto de estudiantes; con los datos obtenidos se ha de construir un banco de ítems calibrado, banco que será utilizado en la implementación de los algoritmos de selección de preguntas abordados en esta investigación. Por último, los resultados arrojados por cada algoritmo serán comparados para determinar cuál es más preciso.

4. Implementación

4.1. Introducción

A continuación, se describirá el diseño de los algoritmos por Conceptos Débiles y por Función de Información, incluyendo los datos de entrada, la construcción del banco de ítems, la implementación de los algoritmos y el entorno de desarrollo.

4.2. Datos de entrada

Los datos base usados en este proyecto fueron obtenidos del estudio realizado por Robles et al. (2012), en el cual se aplica un test que evalúa siete conceptos,como se muestra en la Figura 4.1, a un conjunto de personas, y con los resultados se formó la base de datos mencionada.

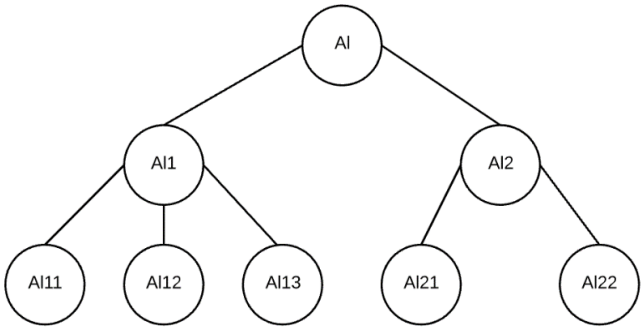

Figura 4.1. Árbol con la estructura temática evaluada en el test. En Descubrimiento de problemas de aprendizaje a través de test: fiabilidad y metodología de diagnóstico basado en clustering por Robles, L., y Rodríguez-Artacho M., 2012.

La Figura 4.2. muestra la estructura de los registros base.

```
S012  A11   A12   A111  A112  A113  A121  A122
S021  A11   A12   A111  A112  A113  A121  A122
S024  A11   A12   A111  A112  A113  A121  A122
S025  A11   A12   A111  A112  A113  A121  A122
S027  A11   A12   A111  A112  A113  A121  A122
S033  A11   A12   A111  A112  A113  A121  A122
                    ...
S345  A11   A12   A111  A112  A113  A121  A122
S348  A11   A12   A111  A112  A113  A121  A122
S351  A12   A111  A121  A122
S354  A11   A12   A111  A113  A121  A122
S357  A12   A111  A113  A121  A122
S358  A11   A12   A111  A112  A113  A121  A122
S360  A11   A12   A111  A112  A113  A121  A122
```

Figura 4.2. Base de datos de conceptos débiles por cada evaluado

4.3. Generación del banco de ítems

Uno de los elementos fundamentales de un TAI es el banco de preguntas, porque de este serán tomados los ítems a presentar a los evaluados. De acuerdo con las características de la presente investigación, y teniendo en cuenta que el propósito específico es seleccionar ítems que evalúen los mismos conceptos que han sido identificados como conceptos débiles en los evaluados, se determinó crear un banco de ítems simulado, que permitiera crear todas las posibles combinaciones de conceptos que evalúan una temática, así como todas las posibles combinaciones de pesos de esos conceptos dentro de cada ítem.

Se siguieron los siguientes pasos para construir el banco de ítems:

- Determinar el número mínimo de ítems del banco.
- Generar tests simulados cuyos ítems hayan sido acertados o desacertados aleatoriamente.
- Medir la fiabilidad y validez de los tests.
- Escoger preguntas de anclaje a partir de los tests válidos.
- Calibrar el número total de ítems en el banco.

Para saber cuántos ítems debía tener el banco de preguntas se tuvo en cuenta que los ítems evaluarían 1 o 2 conceptos fuertemente o muy fuertemente (pesos 4 y 5), y que el total de conceptos a evaluar en un test es de 7 (porque fue circunscrito a una temática que maneja 7 conceptos). Por lo tanto, se calcularon todas las permutaciones posibles de los pesos 4 y 5 en un conjunto de 7 elementos, obteniendo un resultado de 98 permutaciones, es decir, se necesitaría construir un banco que tuviese 98 ítems.

Continuando con el segundo paso, se generaron tests con sus respectivos patrones de respuesta con la función *rmvlogis* del lenguaje de programación R. Esta función genera patrones de respuesta aleatorios según los modelos dicotómicos y politómicos de la TRI, y acepta como primer parámetro el número de patrones de respuesta por test a simular, y como segundo parámetro, una matriz numérica cuyas filas representan los ítems y las columnas, los parámetros. Por lo tanto, la implementación de esta función para generar un test de 20 ítems podría hacerse de la siguiente forma:

```
rmvlogis(10, cbind(seq(-9, 10, 1), 1))
```

La fiabilidad de los tests debía ser asegurada. Por lo tanto, se tuvo en cuenta el alfa de Cronbach de cada test, y sólo fueron considerados fiables los tests cuyo alfa de Cronbach fuese mayor o igual 0.6. La función de R usada para calcular este valor fue la siguiente:

```
cronbach.alpha(k)
```

Donde k representa un conjunto de respuestas generado con la función *rmvlogis*.

Con el objetivo de aplicar el algoritmo de selección de ítems basado en el valor de la *Función de Información*, es necesario que el banco de preguntas a usar esté propiamente calibrado, lo que implica que los valores para los parámetros correspondientes al modelo a implementar sean calculados. Para efectos de la presente investigación, se debía calibrar el banco de ítems según los modelos logísticos de 1 parámetro (Rasch), de 2 parámetros (ltm), y de 3 parámetros (tpm).

En el caso del primer modelo se calculó el valor del discriminante para cada ítem con la siguiente función de R:

```
rasch(k)
```

Esta función retorna una matriz con los coeficientes del nivel de dificultad y el discriminante para un test. Sin embargo, el valor del determinante es siempre el mismo número, pues en el modelo Rasch sólo se tiene en cuenta el nivel de dificultad del ítem.

Para calcular el discriminante y el nivel de dificultad de cada pregunta se usó la función ltm de R, cuya implementación se muestra a continuación:

```
ltm(k ~ z1)
```

Esta función acepta una fórmula como primer parámetro que representa las dependencias entre los ítems de un test. Al lado izquierdo de la fórmula se indica el conjunto de datos (k) y al lado derecho, las variables latentes (z1 o z2) que describen las dependencias antes mencionadas.

Por último, se calcularon el discriminante, el nivel de dificultad y el factor de adivinanza de cada ítem en el banco con la función tpm() de R. Su implementación se muestra a continuación:

```
tpm(k)
```

La siguiente figura muestra como quedó diseñado el banco de ítems para el modelo de 3P:

Ítem	Descripción	Factor de Adivinanza	Factor de Dificultad	Discriminante	Conceptos que evalúa
1	Ítem que evalúa 1 o 2 conceptos	-2.091102	10.9437813	7.475695	[5, 0, 0, 0, 0, 0, 0]
2	Ítem que evalúa 1 o 2 conceptos	-2.144643	21.2071357	30.520184	[0, 5, 0, 0, 0, 0, 0]
3	Ítem que evalúa 1 o 2 conceptos	-22.943176	1.2047660	2.672897	[0, 0, 5, 0, 0, 0, 0]
4	Ítem que evalúa 1 o 2 conceptos	-20.059612	0.8572268	1.811046	[0, 0, 0, 5, 0, 0, 0]
5	Ítem que evalúa 1 o 2 conceptos	-8.497956	3.0387933	1.661660	[0, 0, 0, 0, 5, 0, 0]
6	Ítem que evalúa 1 o 2 conceptos	-3.859219	-3.6801869	47.215701	[0, 0, 0, 0, 0, 5, 0]
7	Ítem que evalúa 1 o 2 conceptos	-4.646742	-28.2795502	44.316990	[0, 0, 0, 0, 0, 0, 5]
8	Ítem que evalúa 1 o 2 conceptos	-3.953342	-17.1289796	27.058213	[4, 0, 0, 0, 0, 0, 0]
9	Ítem que evalúa 1 o 2 conceptos	-17.417162	-2.1413317	4.337192	[0, 4, 0, 0, 0, 0, 0]
10	Ítem que evalúa 1 o 2 conceptos	-12.670661	-2.4112424	2.497437	[0, 0, 4, 0, 0, 0, 0]
11	Ítem que evalúa 1 o 2 conceptos	-29.1637107	2.2460662	0.1303201	[0, 0, 0, 4, 0, 0, 0]
12	Ítem que evalúa 1 o 2 conceptos	1.4781169	-102.9020630	100.6093710	[0, 0, 0, 0, 4, 0, 0]
13	Ítem que evalúa 1 o 2 conceptos	0.8755853	-200.9847745	200.5233545	[0, 0, 0, 0, 0, 4, 0]
14	Ítem que evalúa 1 o 2 conceptos	0.4825535	-1023.7571282	754.4233750	[0, 0, 0, 0, 0, 0, 4]
15	Ítem que evalúa 1 o 2 conceptos	0.2726670	-2960.0756144	2176.3609579	[5, 5, 0, 0, 0, 0, 0]
16	Ítem que evalúa 1 o 2 conceptos	-0.8929050	-2011.3863948	408.7792153	[5, 0, 5, 0, 0, 0, 0]
...					
80	Ítem que evalúa 1 o 2 conceptos	-2.9269538	-2673.4797979	1584.6298702	[4, 0, 0, 5, 0, 0, 0]
81	Ítem que evalúa 1 o 2 conceptos	0.4093778	-43.6258360	136.2204567	[4, 0, 0, 0, 5, 0, 0]
82	Ítem que evalúa 1 o 2 conceptos	-1.3746038	-165.0813970	557.8130248	[4, 0, 0, 0, 0, 5, 0]
83	Ítem que evalúa 1 o 2 conceptos	-3.0439398	-232.8509995	344.8501671	[4, 0, 0, 0, 0, 0, 5]
84	Ítem que evalúa 1 o 2 conceptos	-2.0185519	-389.7100749	455.3074635	[0, 4, 5, 0, 0, 0, 0]
85	Ítem que evalúa 1 o 2 conceptos	1.0314031	-24.9297819	28.6561433	[0, 4, 0, 5, 0, 0, 0]
86	Ítem que evalúa 1 o 2 conceptos	-0.7483697	-80.2576267	58.0474460	[0, 4, 0, 0, 5, 0, 0]
87	Ítem que evalúa 1 o 2 conceptos	-3.3025064	-153.8211456	225.9476724	[0, 4, 0, 0, 0, 5, 0]
88	Ítem que evalúa 1 o 2 conceptos	-77.5457901	-1.6953977	0.4278531	[0, 4, 0, 0, 0, 0, 5]
89	Ítem que evalúa 1 o 2 conceptos	-3.3025087	-154.5279712	226.9901565	[0, 0, 4, 5, 0, 0, 0]
90	Ítem que evalúa 1 o 2 conceptos	-2.0185419	-585.7172009	587.5594809	[0, 0, 4, 0, 5, 0, 0]
91	Ítem que evalúa 1 o 2 conceptos	-37.4874195	0.61695770	2.4301342	[0, 0, 4, 0, 0, 5, 0]
92	Ítem que evalúa 1 o 2 conceptos	-41.9945962	0.28453311	2.5765360	[0, 0, 4, 0, 0, 0, 5]
93	Ítem que evalúa 1 o 2 conceptos	-23.0816922	-0.57871039	0.8940708	[0, 0, 0, 4, 5, 0, 0]
94	Ítem que evalúa 1 o 2 conceptos	-41.9673053	0.14507131	2.2491141	[0, 0, 0, 4, 0, 5, 0]
95	Ítem que evalúa 1 o 2 conceptos	-40.4078560	0.07379790	1.6907808	[0, 0, 0, 4, 0, 0, 5]
96	Ítem que evalúa 1 o 2 conceptos	-1.6272793	-0.09448121	73.4338258	[0, 0, 0, 0, 4, 5, 0]
97	Ítem que evalúa 1 o 2 conceptos	-2.7628225	-0.53883293	3.5954864	[0, 0, 0, 0, 4, 0, 5]
98	Ítem que evalúa 1 o 2 conceptos	-38.0538784	-3.94499405	4.0331040	[0, 0, 0, 0, 0, 4, 5]

Figura 4.3. Matriz de 98 ítems que relaciona los parámetros de cada ítem y los conceptos que evalúa.

4.4. Implementación de los algoritmos de selección de ítems

A continuación, se aplicarán los dos algoritmos de selección de ítems con cada uno de los modelos logísticos tratados en este proyecto.

Algoritmo de Selección de Ítems por Función de Información

- **Datos de entrada**

El algoritmo por Función de Información ha de escoger el primer ítem que el evaluado deberá responder en base a su habilidad en el momento en que inicia un test. Para efectos de esta investigación, la habilidad del alumno corresponde al puntaje que obtuvo previamente en el test de estructura fija. Conociendo el ID del alumno, su puntaje puede ser consultado en los registros base.

- **Detalles de la implementación**

Luego de obtener el puntaje del evaluado, se ha de calcular la información que cada ítem en el banco proporciona. Este valor se obtiene aplicando las fórmulas mencionadas en el Capítulo 3, Fórmulas III.5 y III.6, las cuales depende del modelo logístico.

La primera pregunta del test ha de corresponder con el ítem que mayor información aporte. Por lo tanto, de la estructura de datos presentada arriba se escogió la clave cuyo valor fuese el máximo. Es entonces cuando la primera pregunta, que corresponde con la clave seleccionada, dio comienzo al test y aleatoriamente se dio una respuesta a ésta. Para escoger el siguiente ítem de un test, se hizo el cálculo de la habilidad utilizando la fórmula mencionada en el Capítulo 3, Fórmula III.4, dependiendo del modelo implementado.

El criterio de parada en esta simulación fue un máximo de ítems presentados igual a 8.

- **Datos de salida –ítems seleccionados**

La ejecución de este algoritmo arrojó un conjunto de preguntas seleccionadas y una matriz donde se relacionan los conceptos débiles que evalúa cada ítem.

Algoritmo de Selección de Ítems por Conceptos Débiles

- **Datos de entrada**

El algoritmo por Conceptos Débiles ha de generar un pretest con los ítems a ser presentados al evaluado en base a sus deficiencias conceptuales. Por lo tanto, luego de obtener el ID del alumno, sus conceptos débiles han de ser buscados en los registros base.

- **Detalles de la implementación**

Con los conceptos débiles del estudiante y usando CBR se escogieron los ítems que evaluarían esos mismos conceptos. La función que genera este pretest se llama `find_suitable_questions` y se encuentra en el archivo `pretest.py`. Esta función recibe una lista de conceptos débiles y crea grupos o casos en base a éstos; por cada concepto crea una lista de 0 y 1, y en la posición correspondiente al concepto en cada iteración coloca un 1, mientras que en los demás coloca 0. Se obtuvieron tantos casos como conceptos débiles. Por ejemplo, para el concepto Al1 se crea el grupo [1, 0, 0, 0, 0, 0, 0]. Luego, esta función recorre el archivo `questions-weights.csv` calculando la distancia euclidiana entre los casos y los conceptos que evalúa cada ítem, y escoge 2 ítems, por cada concepto débil, cuyas distancias euclidianas sean las menores.

Los 2 ítems escogidos se guardan en un diccionario junto con el concepto débil que evalúan. Este diccionario es el pretest. Entonces, se da inicio al test simulando que la primera pregunta del pretest es presentada a un evaluado. Después de tener una respuesta para cada ítem, se calcula vez tras vez la aptitud del alumno.

- **Datos de salida – ítems seleccionados**

La ejecución de este algoritmo arrojó un conjunto de preguntas seleccionadas y una matriz donde se relacionan los conceptos débiles que evalúa cada ítem.

4.5. Descripción del entorno de desarrollo

a. Python.

Python es un lenguaje de programación creado por Guido van Rossum a finales de los ochenta, y que gracias a sus características ha llegado a ser un lenguaje muy conocido en la actualidad.Es un lenguaje muy simple, por lo que es muy fácil iniciarse en este lenguaje. El pseudo-código natural de Python es una de sus grandes fortalezas. Usando el lenguaje Python se puede crear todo tipo de programas; programas de propósito general y también se pueden desarrollar páginas Web.

En esta investigación, se trabajó con la versión de Python 3.5.2 en el entorno PyCharm Professional Edition versión 2017.1.4.

b. R Project.

R es un lenguaje de programación interpretado, y se mantiene en un ambiente para el cómputo estadístico y gráfico. El término ambiente pretende caracterizarlo como un sistema totalmente planificado y coherente, en lugar de una acumulación gradual de herramientas muy específicas y poco flexibles, como suele ser con otro software de análisis de datos. En lugar de pensar de R como un sistema estadístico, es preferible verlo como un ambiente en el que se aplican técnicas estadísticas.

En esta investigación, se trabajó con la versión de R3.3.2en el entorno RStudio versión 1.0.136.

4.6. Conclusiones

Haber descrito el proceso de implementación permitió demostrar los pasos seguidos para llevar a la práctica la metodología trazada y aclarar dudas sobre cómo se lograrán los objetivos de este proyecto. Asimismo, se detallaron las herramientas usadas equipando a futuros autores con el conocimiento generado en esta investigación.

5. Pruebas

5.1. Introducción

En este capítulo se mostrarán los resultados arrojados por el programa desarrollado, es decir, los ítems que cada algoritmo escogió del banco, y se compararán con los conceptos débiles del evaluado para medir cuán preciso es cada algoritmo en la selección de ítems.

5.2. Prueba del Algoritmo de Selección por Conceptos Débiles

El perfil escogido para hacer las pruebas fue el del evaluado con ID S075:

ID	C1	C2	C3	C4	C5	C6	C7	SCORE
S075	Al22	Al2	Al11	Al21	0	0	0	46.1

Figura 5.1. Perfil de entrada del evaluado S075.

Modelo de Rasch

En base a sus conceptos débiles se generó el siguiente pretest:

```
Pretest {6: [59, 77], 4: [30, 48], 1: [45, 64], 5: [28, 40]}
```

Figura 5.2. Pretest generado para el estudiante S075 según el modelo de Rasch.

Los números 1, 4, 5 y 6 representan los conceptos Al11, Al2, Al21 y Al22, respectivamente, y su valor correspondiente se traduce en los ítems que le serán presentados al estudiante y que evalúan tal concepto, es decir, los ítems 45 y 64 evalúan el concepto Al11; los ítems 30 y 48 evalúan el concepto Al2; los ítems 28 y 40 evalúan el concepto Al21, y los ítems 59 y 77 evalúan el concepto Al22.

Entonces se simuló la respuesta del estudiante a cada uno de estos ítems. Después de que cada pregunta fue respondida, se estimó la habilidad del estudiante generando una matriz como la que se muestra en la Figura 5.3., donde cada columna es un número del 0 al 100, que es la escala manejada en esta investigación para el cálculo de la habiliad, y luego se aplicó la fórmula mencionada en el Capítulo 3, Fórmula III.4 para obtener la estimación de la aptitud.

```
        0         1         2         3         4         5         6    \
0.836512  0.965521  0.993518  0.998809  0.999782  0.99996  0.999993

        7   8   9   ... 91  92  93  94  95  96  97  98  99  100
0.999999   1   1 ...   1   1   1   1   1   1   1   1   1   1
```

Figura 5.3. Matriz de probabilidades según el modelo de Rasch

Luego de repetir estos pasos para cada ítem en el pretest, se observaron las preguntas escogidas por el algoritmo.

Item	A11	A111	A112	A113	A12	A121	A122
28	0	0	1	0	0	1	0
30	0	0	0	1	1	0	0
40	1	0	0	0	0	1	0
45	0	1	0	0	0	1	0
48	0	0	1	0	1	0	0
59	1	0	0	1	0	0	0
64	0	1	0	1	0	0	0
77	0	0	0	0	0	1	1

Figura 5.4. Conceptos que evalúa cada ítem en el pretest

Modelo ltm

En base a los conceptos débiles del evaluado se generó el siguiente pretest:

```
Pretest {1: [21, 43], 4: [5, 55], 5: [40, 52], 6: [53, 83]}
```

Figura 5.5. Pretest generado para el estudiante S075 según el modelo ltm.

Después de que cada pregunta fue respondida, se estimó la habilidad del estudiante generando una matriz como la que se muestra en la Figura 5.6. y luego se aplicó la fórmula de la EMV para calcular la habilidad.

```
        0         1         2         3         4         5         6    \
0.984632  0.996214  0.999075  0.999775  0.999945  0.999987  0.999997

        7   8   9   ... 91  92  93  94  95  96  97  98  99  100
0.999999   1   1 ...   1   1   1   1   1   1   1   1   1   1
```

Figura 5.6. Matriz de probabilidades según el modelo ltm

Luego de repetir estos pasos para cada ítem en el pretest, se observaron las preguntas escogidas por el algoritmo.

Item	A11	A111	A112	A113	A12	A121	A122
5	0	0	0	0	1	0	0
21	0	1	1	0	0	0	0
40	1	0	0	0	0	1	0
43	0	1	0	1	0	0	0
52	0	0	0	1	0	1	0
53	0	0	0	1	0	0	1
55	0	0	0	0	1	0	1
83	1	0	0	0	0	0	1

Figura 5.7. Conceptos que evalúa cada ítem en el pretest

Modelo tpm

En base a los conceptos débiles del evaluado se generó el siguiente pretest:

```
Pretest {6: [20, 59], 4: [51, 93], 1: [22, 42], 5: [19, 94]}
```

Figura 5.8. Pretest generado para el estudiante S075 según el modelo tpm.

Después de que cada pregunta fue respondida, se estimó la habilidad del estudiante generando una matriz como la que se muestra en la Figura 5.6. y luego se aplicó la fórmula de la EMV para calcular la habilidad.

0	1	2	3	4	5	6	7	\
1	1	1	1	1	1	1	1	
1	1	1	1	1	1	1	1	
-30.0786	-27.465	-20.0884	-9.1489	-2.21403	0.163491	0.794816	0.950429	
1	1	1	1	1	1	1	1	
-9.44266	-8.12189	-6.74593	-5.39182	-4.13183	-3.01911	-2.0809	-1.32021	
1	1	1	1	1	1	1	1	
-3.42892	-2.9993	-2.55093	-2.0987	-1.65805	-1.24286	-0.863886	-0.527833	
1	1	1	1	1	1	1	1	

8	9	...	91	92	93	94	95	96	97	98	99	100
1	1	...	1	1	1	1	1	1	1	1	1	1
1	1	...	1	1	1	1	1	1	1	1	1	1
0.988068	0.997131	...	1	1	1	1	1	1	1	1	1	1
1	1	...	1	1	1	1	1	1	1	1	1	1
-0.72279	-0.265242	...	1	1	1	1	1	1	1	1	1	1
1	1	...	1	1	1	1	1	1	1	1	1	1
-0.237416	0.00803353	...	1	1	1	1	1	1	1	1	1	1
1	1	...	1	1	1	1	1	1	1	1	1	1

Figura 5.9. Matriz de probabilidades según el modelo tpm

Luego de repetir estos pasos para cada ítem en el pretest, se observaron las preguntas escogidas por el algoritmo.

Item	A11	A111	A112	A113	A12	A121	A122
19	1	0	0	0	0	1	0
20	1	0	0	0	0	0	1
22	0	1	0	1	0	0	0
42	0	1	1	0	0	0	0
51	0	0	0	1	1	0	0
59	1	0	0	1	0	0	0
93	0	0	0	1	1	0	0
94	0	0	0	1	0	1	0

Figura 5.10. Conceptos que evalúa cada ítem en el pretest

5.3. Prueba del Algoritmo de Selección por Función de Información

Para simular este algoritmo fue necesario obtener primero el puntaje del alumno mostrado en la Figura 5.1. Con este puntaje se escogió la primera pregunta calculando la Función de Información de todos los ítems en el banco de preguntas y escogiendo aquel que aportara mayor información. Para lograr este objetivo se guardó en un diccionario, como se muestra en la Figura 5.8., cada valor resultante junto con el ítem correspondiente.

Modelo de Rasch

Para el evaluado se generó un diccionario con los valores de la función de información de cada ítem:

```
{2: 0.0045479441263225149, 3: 0.066362796442948457, 4: 0.085731283036054706, 5: 0.00026942811470686412,
```

Figura 5.11. Diccionario con la IIF de cada ítem según el modelo de Rasch

Luego, se seleccionó el ítem con mayor IIF y se dio comienzo al test. Después de dar respuesta a un ítem se recalculó el nivel de aptitud para nuevamente hacer el procedimiento descrito arriba, pero esta vez se generó la matriz de probabilidades de acierto de cada ítem según el modelo de Rasch como se muestra en la siguiente figura:

0	1	2	3	4	5	\
8.86775e-05	0.000485138	0.0026494	0.0143303	0.0737051	0.303371	
0.00456882	0.0245043	0.120864	0.429363	0.804612	0.957515	
0.0714709	0.296403	0.697482	0.92657	0.985727	0.997361	
0.0946992	0.364071	0.758062	0.944899	0.989457	0.998057	
0.000269501	0.0014732	0.00801	0.0423222	0.19476	0.569656	
0.00770742	0.0407768	0.188745	0.560117	0.874513	0.974451	
0.00984085	0.051588	0.229405	0.61967	0.899164	0.979921	
0.00984085	0.051588	0.229405	0.61967	0.899164	0.979921	
0.00682793	0.0362617	0.170762	0.529862	0.860496	0.97123	
0.00692544	0.036764	0.172794	0.533417	0.862201	0.971626	

6	7	8	9	...	91	92	93	94	95	96	97	98	\
0.704439	0.928797	0.986186	0.997447	...	1	1	1	1	1	1	1	1	
0.991958	0.998521	0.999729	0.999951	...	1	1	1	1	1	1	1	1	
0.999517	0.999912	0.999984	0.999997	...	1	1	1	1	1	1	1	1	
0.999644	0.999935	0.999988	0.999998	...	1	1	1	1	1	1	1	1	
0.87871	0.9754	0.995413	0.999159	...	1	1	1	1	1	1	1	1	
0.995232	0.999125	0.99984	0.999971	...	1	1	1	1	1	1	1	1	
0.99627	0.999316	0.999875	0.999977	...	1	1	1	1	1	1	1	1	
0.99627	0.999316	0.999875	0.999977	...	1	1	1	1	1	1	1	1	
0.994617	0.999012	0.999819	0.999967	...	1	1	1	1	1	1	1	1	
0.994693	0.999026	0.999822	0.999967	...	1	1	1	1	1	1	1	1	

Figura 5.12. Matriz de probabilidades de acierto de los ítems según el modelo de Rasch

Al finalizar la simulación con este método se escogieron las siguientes preguntas:

Item	A11	A111	A112	A113	A12	A121	A122
1	1	0	0	0	0	0	0
22	0	1	0	1	0	0	0
32	0	0	0	1	0	0	1
55	0	0	0	0	1	0	1
61	1	0	0	0	0	1	0
71	0	0	1	0	0	0	1
85	0	1	0	1	0	0	0
97	0	0	0	0	1	0	1

Figura 5.13. Preguntas seleccionadas por el algoritmo por IIF y el modelo de Rasch

Modelo ltm

Siguiendo el mismo procedimiento descrito en el inciso anterior, se escogieron las siguientes preguntas:

Item	A11	A111	A112	A113	A12	A121	A122
6	0	0	0	0	0	1	0
20	1	0	0	0	0	0	1
82	1	0	0	0	0	1	0
83	1	0	0	0	0	0	1
87	0	1	0	0	0	1	0
89	0	0	1	1	0	0	0
92	0	0	1	0	0	0	1
97	0	0	0	0	1	0	1

Figura 5.14. Preguntas seleccionadas por el algoritmo por IIF y el modelo ltm

Modelo tpm

Al implementar este modelo se escogieron las preguntas mostradas en la Figura siguiente:

Item	A11	A111	A112	A113	A12	A121	A122
11	0	0	0	1	0	0	0
16	1	0	1	0	0	0	0
17	1	0	0	1	0	0	0
41	1	0	0	0	0	0	1
42	0	1	1	0	0	0	0
47	0	0	1	1	0	0	0
58	1	0	1	0	0	0	0
76	0	0	0	0	1	0	1

Figura 5.15. Preguntas seleccionadas por el algoritmo por IIF y el modelo tpm

5.4. Análisis de Resultados

Los ítems escogidos por cada algoritmo usando los distintos modelos fueron comparados para determinar qué algoritmo fue más preciso al presentar ítems que evaluaran justamente los conceptos débiles de un estudiante. Se presentan a continuación comparaciones hechas por cada modelo implementado entre los dos algoritmos. En cada tabla se relacionan el número de ítems presentados que evaluaran los conceptos débiles del estudiante.

Modelo de Rasch

Concepto	Algoritmo por Conceptos Débiles	Algoritmo por Función de Información
Al22	1	4
Al2	2	2
Al11	2	2
Al21	4	1

Tabla 5.1. Relación entre los ítems presentados por cada concepto usando el modelo de Rasch

Según la Tabla 5.1., ambos algoritmos tuvieron un desempeño parecido en términos de precisión. Ambos evaluaron todos los conceptos presentados, pero algunos más fuertemente que otros.

Modelo ltm

Concepto	Algoritmo por Conceptos Débiles	Algoritmo por Función de Información
Al22	3	4
Al2	2	1
Al11	2	1
Al21	2	3

Tabla 5.2. Relación entre los ítems presentados por cada concepto usando el modelo ltm

La tabla muestra que el algoritmo por Conceptos Débiles presentó los ítems de manera más o menos balanceada, mientras que el algoritmo por Función de Información escogió sólo 1 ítem para evaluar 2 conceptos y, 4 y 3 para evaluar otros.

Modelo tpm

Concepto	Algoritmo por Conceptos Débiles	Algoritmo por Función de Información
Al22	1	2
Al2	2	1
Al11	2	1
Al21	2	0

Tabla 5.3. Relación entre los ítems presentados por cada concepto usando el modelo tpm

La relación entre los algoritmos usando el modelo tpm demuestra que el algoritmo por Conceptos Débiles tuvo mayor precisión al seleccionar ítems. Ningún concepto dejó de ser evaluado, mientras que el algoritmo por Función de Información falló en evaluar 1 concepto.

6. Conclusiones y Trabajos Futuros

6.1. Conclusiones

A continuación, se anotan las conclusiones extraídas de éste trabajo:

- Del análisis de resultados hecho se puede concluir que el algoritmo de Selección de Ítems por Conceptos Débiles es más preciso a la hora de escoger preguntas que estén relacionadas justamente con las deficiencias que un test busca evaluar, mientras que el algoritmo por Función de Información no garantiza que las falencias de un alumno sean evaluadas, lo cual supondría una falla en el objetivo principal de una evaluación. Por lo tanto, se recomienda continuar el desarrollo de este algoritmo como se mostrará en el siguiente apartado.

6.2. Trabajos Futuros

A continuación, se recomiendan algunas líneas de investigación que nacen de éste proyecto:

- El algoritmo de Selección de Ítems por Conceptos Débiles fue implementado con un pretest que incluía los ítems que evaluaban las deficiencias conceptuales del alumno. En el futuro podría implementarse éste algoritmo sin crear un pretest, sino escogiendo ítems dependiendo de la respuesta del alumno a cada pregunta.
- Una interfaz de usuario que permitiera observar claramente la comparación entre los dos algoritmos estudiados en este proyecto sería útil para quienes estén interesados en la temática.
- Se podría investigar sobre otros algoritmos de selección de ítems que hayan sido ampliamente usados y contrastar su precisión contra la del algoritmo por Conceptos Débiles.

Referencias Bibliográficas

Anohina, A., Graudina, V., Grundspenkis, J. "Using Concept Maps in Adaptive Knowledge Assessment", Advances in Information Systems Development, Springer US, pp. 469 – 479. 2007.

Boticario Jesús González, y Guadioso V. Elena. Aprender y Formar en Internet. Madrid, España. Thomson Learning Paraninfo. 2001.

Boticario Jesús González, y Guadioso V. Elena. Sistemas Interactivos de Enseñanza/Aprendizaje. Madrid, España. Sanz y Torres. 2003.

Chen, S. & Zhang, J. (2008). *Ability Assessment Based on CAT in Adaptive Learning System.* En2008 International Workshop On Education Technology And Training & 2008 International Workshop On Geoscience And Remote Sensing.

Gronlund, N. E. Elaboración de test de aprovechamiento. México: Trillas. 1976.

Gutiérrez, J., Pérez, T. A., López-Cuadrado, J., Arrubarrena, R. M. y Vadillo, J. A. Evaluación en sistemas hipermedia adaptativos. *Revista de Metodología de las Ciencias del Comportamiento*, Volumen especial, 279-283. 2002.

Guzmán, E. &Conejo, R. (2005). *Self-Assessment in a Feasible, Adaptive Web-Based Testing System.* EnIEEE Transactions On Education, 48(4) (pp. 688-695).

Huapaya, C., Lizarralde, F., Vivas, J., Arona, G. Modelo de Evaluación del Conocimiento en un Sistema Tutorial Inteligente. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, No. 2. 2007.

Hambleton, R.K. & Rovinelli, R.J. Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302. 1986.

Hammond, N. Learning with hypertext: Problems principles and prospects. En: Hypertext a psychological perspective, págs. 51-70. Ellis Horwood (Nueva York), 1993.

Jadhav, M., Rizwan, S. &Nehete, A. (2013). *User Profiling based Adaptive Test Generation and Assessment in E–Learning System.* En 2013 3rd IEEE International Advance Computing Conference.

Lee, Y., Cho, J., Han, S., & Choi, B. (2010).*A Personalized Assessment System based on Item Response Theory.* En International Conference on Web-based Learning 2010, Lecture Notes in Computer Science 6483 (pp. 381–386).

Lin, C., Chen, K., & Tsai, C. (2008). *Modeling the Examinee Ability on the Computerized Adaptive Testing Using Adaptive Network-Based Fuzzy Inference System.* En 2008 IEEE Asia-PacificServices Computing Conference.

Muñiz Fernández J. Introducción a la Teoría de Respuesta a los Items. Madrid: Pirámide. 1997.

Muñiz, J. (2010). *Las Teorías de los Tests: Teoría Clásica y Teoría de Respuesta a los Ítems.* En Papeles del Psicólogo, 31(1) (pp. 57-66).

Rasch, G. Probabilistic models for some intelligence and attainment test. Danish Institute for Educational Research. 1960.

Robles, L., & Rodríguez M. (2012). *Descubrimiento de problemas de aprendizaje a través de test: fiabilidad y metodología de diagnóstico basado en clustering.* (Tesis de maestría). Universidad Nacional de Educación a Distancia - UNED, Madrid, España.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101. 1904.

Spearman, C. Demonstration of formulae for true measurement of correlation. American Journal of Psychology, 18, 161-169. 1907.

Spearman, C. Correlations of sums and differences. British Journal of Psychology, 5, 417-426. 1913.

VanLehn, K. y Martin, J. Evaluation of an assessment system based on Bayesian Student Modeling. *International Journal of Artificial Intelligence and Education*, 8 (2). 1998.

Wainer, H. Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum. 1990.

Yela M. Los test y el análisis factorial. Psicothema. 8(sup):73-88. 1996.