

Sistema de recomendación vocacional en línea

Online vocational recommendation system

Sergio Alcaraz Corona

Universidad Autónoma de Nuevo León, México

sergio.alcarazcrn@uanl.edu.mx

Mario Alejandro Bustos Jiménez

Universidad Autónoma de Nuevo León, México

mariobustosjmz@gmail.com

Resumen

Actualmente elegir la profesión deseada se ha convertido en una de las decisiones más difíciles dada la gran variedad de opciones e información disponibles. El proceso para seleccionar una profesión comienza con la correcta elección del programa educativo que más se ajuste a las habilidades, conocimiento y preferencias del individuo. Para apoyar dicho proceso existen diferentes opciones tales como cursos de orientación vocacional así como aplicaciones informáticas, las cuales ayudan a los usuarios en su proceso de selección (katz, 1993); sin embargo, la mayoría de estas aplicaciones tienen un costo alto o no ofrecen resultados confiables. El objetivo de este trabajo es presentar un sistema de recomendación vocacional en línea basado en un algoritmo de asociación y clasificación de información como el principal componente de definición de perfiles vocacionales.

Palabras clave: sistemas de recomendación, orientación vocacional, inteligencia artificial.

Abstract

Now choose the desired profession has become one of the most difficult decisions given the large variety of available options and information. The process to select a profession begins with the right choice of the educational program that suits the skills, knowledge and preferences of the individual. To support this process, there are different options such as courses of vocational guidance as well as computer applications, which help the users in their selection process (katz, 1993); however, most of these applications have a high cost or do not provide reliable results. The objective of this work is to present a recommendation vocational online system based on an algorithm of Association and classification of information as the main component of vocational profiles.

Key words: recommendation, vocational orientation, artificial intelligence systems.

Fecha recepción: Mayo 2015

Fecha aceptación: Noviembre 2015

Introducción

Una de las decisiones más importantes en la vida de cualquier persona es seleccionar su profesión de manera adecuada. La dificultad para tomar dicha decisión ha aumentado debido a diversos factores, principalmente económicos, tecnológicos y sociales. En un mundo que cambia y avanza cada vez más rápido, la diversidad de opciones profesionales se ha vuelto más amplia y dinámica debido al surgimiento de nuevas profesiones.

Recientemente se han hecho grandes cambios en los planes de estudio de las diversas carreras profesionales que ofrecen casi todas las universidades del mundo, con la finalidad de producir profesionales mejor preparados para un entorno cada vez más competitivo y globalizado. Al mismo tiempo, nuevas carreras profesionales han estado surgiendo para satisfacer nuevas demandas y requerimientos laborales. Lo anterior, junto con otros factores, constituye una gran

cantidad de información que puede hacer que el proceso de selección de una carrera profesional sea una tarea complicada, sobre todo si se carece de elementos o herramientas de apoyo.

Una de las opciones que hay para apoyar el proceso de selección de estudios superiores son los tradicionales cursos de orientación vocacional, elaborados principalmente por instituciones educativas públicas y privadas en diferentes formatos y contenidos. Evidentemente, uno de los componentes esenciales de dichos cursos de orientación vocacional es el cuestionario con preguntas encaminadas a proporcionar al usuario información que le sea útil para su proceso de elección de carrera. Actualmente es muy común encontrar dichos cuestionarios así como otra información, mediante una aplicación o búsqueda de información en internet. Sin embargo, la mayoría de este tipo de cuestionarios y guías son muy costosos o no producen resultados confiables (Franca, 2012). Por eso es indispensable contar con herramientas de apoyo que proporcionen nuevos y mejores servicios de orientación vocacional (Katz, 1993). El auge de las tecnologías de información, principalmente todo lo relacionado con internet, ha contribuido al avance y mejoramiento de una gran cantidad de aplicaciones de la vida diaria.

Actualmente las denominadas aplicaciones web pueden cubrir una amplia gama de necesidades, las cuales van desde adquirir diversos artículos, hacer una reservación o consultar información general, entre otras cosas. En cierto tipo de aplicaciones web, el resultado de una consulta puede ir acompañado de una o más recomendaciones que proporcionen al usuario mayor información para tomar una mejor decisión. Por ejemplo, el portal de compras por internet Amazon.com recomienda cierta cantidad de artículos similares al buscado. Dichas recomendaciones se obtienen por medio de algoritmos especiales empleados en aplicaciones denominadas de inteligencia artificial y minería de datos, las cuales procesan la información que va proporcionando el usuario. Existe una amplia variedad de algoritmos y metodologías para este propósito, sin embargo, un tipo de estos algoritmos busca asociar y clasificar información relacionada entre sí de acuerdo a ciertos criterios o reglas que permiten facilitar el uso y aprovechamiento de dicha información.

El objetivo de este trabajo es diseñar e implementar un sistema de recomendación vocacional en línea empleando lenguajes de programación y tecnologías de información recientes con las que el sistema pueda ser capaz de producir resultados confiables y a la vez generar las mejores recomendaciones de estudio para cada usuario de la aplicación. Por tanto, el sistema de

recomendación propuesto en este proyecto es un sitio web de orientación vocacional para uso principalmente de la comunidad estudiantil de la Facultad de Ingeniería Mecánica y Eléctrica (FIME) de la Universidad Autónoma de Nuevo León (UANL), cuyo identificador URL es <http://www.fime.vocacionaluanl.com>. Para ello, el sistema tomará como base las respuestas que corresponden a las preguntas relacionadas con las capacidades, aptitudes, intereses y gustos que registra el usuario, para que sean todas procesadas y clasificadas por la lógica del sistema y de esa manera determinar en cada caso el perfil vocacional más apropiado.

DESARROLLO

Antecedentes

Tal como su nombre indica, el objetivo principal de cualquier sistema de recomendación es proporcionar cierta información de interés en respuesta a una determinada solicitud o búsqueda por parte del usuario o usuarios (Adomavicius, 2005). Los sistemas de recomendación forman parte de una extensa área de las ciencias computacionales que se conoce comúnmente como inteligencia artificial, cuyo objetivo principal es la gestión y administración del conocimiento e información por medio de sistemas o aplicaciones con cierto nivel de inteligencia para resolver problemas y tomar decisiones dentro de cierto contexto (Truemper, 2004). Entre los sistemas inteligentes más comunes se encuentran los sistemas expertos (Truemper, 2004), las redes neuronales (Kasabov, 1996), la minería de datos (Witten & Frank, 2005) y la lógica difusa (Kasabov, 1996), entre otros.

De cierta forma, los sistemas de recomendación se consideran como un tipo especial de sistema experto. Adicionalmente existe una gran cantidad de aplicaciones con las que los sistemas de recomendación han contribuido a mejorar los procesos de búsqueda y selección de información. Una de las aplicaciones que más auge ha tenido por los beneficios que ofrece tanto a sus usuarios como a las organizaciones que la implementa es la de comercio electrónico, principalmente para la adquisición de productos y servicios comerciales (Leavitt, 2006; Linden et al., 2003).

Aunque no ha sido tan amplio como en el ámbito comercial, se han encontrado algunas aplicaciones de sistemas de recomendación para entornos académicos o educativos (Aher y Lobo, 2013; Sobecki, 2014). En particular, uno de estos trabajos presenta una aplicación de este tipo de sistemas, el cual también utiliza una implementación del algoritmo k means como parte de su

mecanismo para recomendar cursos a estudiantes bajo ciertos criterios o condiciones especiales (Aher y Lobo, 2013).

Una de las características distintivas de los sistemas de recomendación se refiere a la lógica que sigue el procedimiento que se encarga de producir sus recomendaciones y de la cual se derivan las siguientes categorías (Adomavicius, 2005):

Recomendaciones basadas en contenido. Es la categoría que requiere conocer las características relacionadas con el contenido o descripción del o de los elementos que han sido previamente recomendados al momento de producir nuevas recomendaciones para un determinado usuario o grupo de usuarios.

Recomendaciones basadas en filtrado colaborativo. Es otra categoría donde solo se consideran las recomendaciones previamente realizadas a usuarios con gustos, intereses y/o características similares para generar nuevas recomendaciones para un determinado usuario o grupo de usuarios, es decir, en estos casos el contenido o descripción de los elementos recomendados no es relevante.

Los esquemas híbridos son aquellos que combinan funcionalidad de las dos categorías anteriores. En este caso, la lógica del sistema de recomendación propuesto es del tipo de la primera categoría puesto que analiza el contenido; en específico, el valor de la respuesta seleccionada para cada pregunta como parte del proceso de generación de recomendaciones.

Definición de preguntas

El proceso de definición de preguntas comenzó con una revisión de literatura para entender los aspectos más importantes en la elaboración de preguntas que corresponden a un cuestionario de orientación vocacional. De acuerdo a lo consultado, se pudo comprobar que existen varios factores que influyen en la definición de preguntas en este tipo de cuestionarios, tales como habilidades, conocimientos, preferencias, personalidad e intereses de cada individuo, que son los más importantes. En general, los factores anteriores han sido ampliamente estudiados para formar la llamada teoría de Holland de orientación vocacional, la cual ha encontrado, entre otros aspectos, que la personalidad y las habilidades juegan un papel muy importante en cualquier proceso de selección de estudios superiores y que se refieren a lo que un individuo es capaz de hacer en un futuro (Betz et al., 1989; Gati et al., 2006; Holland, 1997).

Por otro lado, también se creó una encuesta en línea (Alcaraz et al., 2015), la cual se envió a grupos de estudiantes universitarios para tratar de identificar los factores que más influencia tuvieron en su proceso de selección de una carrera profesional, así como diversas situaciones que los llevaron a hacer cambios de carrera. Con la aplicación de esta encuesta se obtuvo una muestra de aproximadamente 200, de la que se pudo obtener datos valiosos para la definición del sistema y de esa forma contar con más elementos para desarrollar una herramienta de orientación vocacional más eficiente y confiable.

A partir de la revisión de literatura y los resultados de la encuesta, se procedió a la creación de 50 preguntas para cada carrera administrada por el sistema, que en su versión inicial incluye tres carreras de la Facultad de Ingeniería Mecánica y Eléctrica de la Universidad Autónoma de Nuevo León: Ingeniero en Tecnología de Software (ITS), Ingeniero Administrador de Sistemas (IAS), Ingeniero en Electrónica y Comunicaciones (IEC). Después de analizar los conceptos más relevantes para el proceso de definición, se determinó que la cantidad anterior de preguntas permite representar las características y atributos más importantes de cada carrera de forma eficaz.

Las preguntas se organizan dentro de archivos Excel, comenzando por las más generales hasta las más específicas. Una vez que las preguntas fueron ordenadas, se planificaron niveles (tiers) para las mismas. Los niveles comprenden del 0 al 3, siendo 3 el nivel más representativo de cada carrera mientras que 0 es el nivel más básico o general (tronco común). La distribución aproximada de preguntas que se propuso por nivel es: (nivel, cantidad de preguntas) (0, 10), (1, 15), (2, 15), (3, 10).

La distribución anterior puede variar dependiendo de varios factores, entre los cuales están la antigüedad, el historial, la documentación y la demanda de una carrera en particular. Sin embargo, al organizar las preguntas de la forma anterior se evita tener que implementar un proceso adicional que las clasifique en grupos de acuerdo al grado o nivel de representación que tengan para una determinada carrera. De lo contrario, hubiera sido necesario implementar el algoritmo *k* means (*k* medias), que es capaz de clasificar y agrupar todo un conjunto de *n* observaciones, asignando cada una a uno de *k* grupos posibles (Duda & Hart, 1973). Por lo tanto, el contenido de cada archivo Excel de preguntas se almacena en una base de datos siguiendo el ordenamiento mencionado.

Estructura del sistema de recomendación

Dentro de la cada vez más extensa área de desarrollo de sistemas de información, existe una gran variedad de opciones, lenguajes y herramientas tecnológicas para construir este tipo de aplicaciones web o páginas de internet dinámicas. Para este proyecto se utilizó principalmente PHP (PHP Hypertext Preprocessor), que es un lenguaje de programación de propósito general normalmente empleado para el desarrollo de aplicaciones web con contenido dinámico en arquitecturas cliente-servidor (Beati, 2011). El lenguaje PHP ofrece varias ventajas sobre otros lenguajes de programación que se utilizan para desarrollar aplicaciones del tipo buscado en este proyecto, es decir, una aplicación web dinámica con una base de datos. En la figura 1 se muestra el diagrama de los componentes de software que forman el sistema y se encargan de su operación.

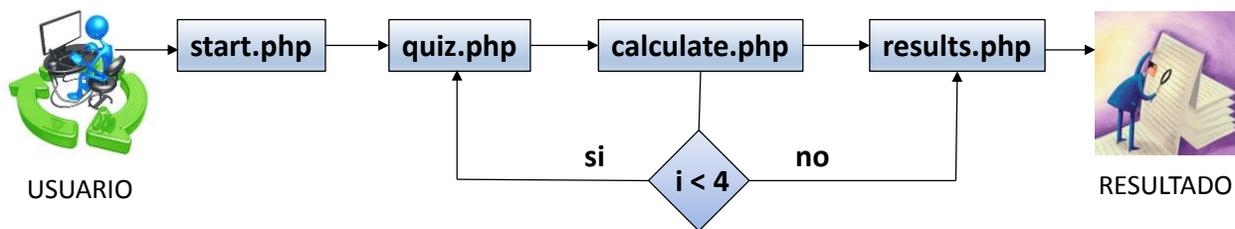


Figura 1. Estructura del sistema de recomendación fime.vocacionaluanl

El módulo `start.php` se encarga de establecer la conexión a la base de datos así como de inicializar en cero la variable de iteración (i), con la que se controla posteriormente la ejecución de los módulos subsecuentes. Una vez realizada la función anterior, se ejecuta el módulo `quiz.php` cuya función principal es desplegar las preguntas del cuestionario comenzando por el nivel más general (nivel 0) para que sean contestadas por el usuario en bloques de 15 por pantalla, como muestra la figura 2. Nótese que las preguntas son todas de opción múltiple y que emplean una escala tipo Likert (Likert, 1932) del 1 al 5, donde 1 representa el valor mínimo de asociación o afinidad para una determinada pregunta mientras que 5 representa el valor máximo. En este caso, la variable i representa el nivel al que pertenecen las preguntas que se van mostrar en un momento determinado.

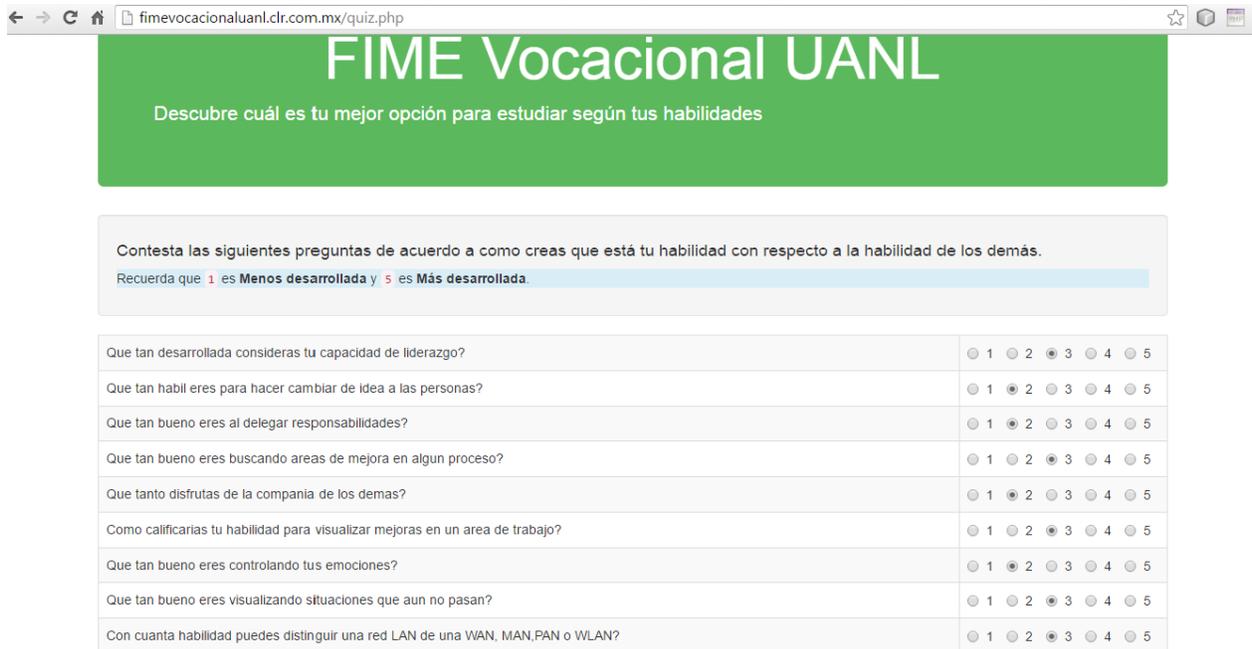


Figura 2. Funcionamiento del sistema de recomendación

Adicionalmente, las preguntas que se van mostrando se extraen de la base de datos del sistema, la cual consta de dos tablas, carreras y preguntas. La figura 3 muestra el diagrama de la base de datos (fime.vocacionaluanlBD) que utiliza el sistema. Por un lado, la tabla carreras almacena los datos correspondientes a las carreras que soporta el sistema comenzando por un identificador único para cada carrera, así como el nombre y la fecha de alta de cada carrera. En este caso, la primera carrera (idcarrera = 1) en ser almacenada en la base de datos fue IAS, seguida de IEC y por último ITS. Por otro lado, la tabla de preguntas contiene todos los enunciados de las preguntas así como su respectivo nivel. La base de datos fime.vocacionaluanlBD se diseñó e implementó con MySQL el cual es un sistema de administración de bases de datos de software libre o abierto que comúnmente se utiliza en el desarrollo de aplicaciones web debido a su funcionalidad, flexibilidad y confiabilidad (MySQL, 2016).

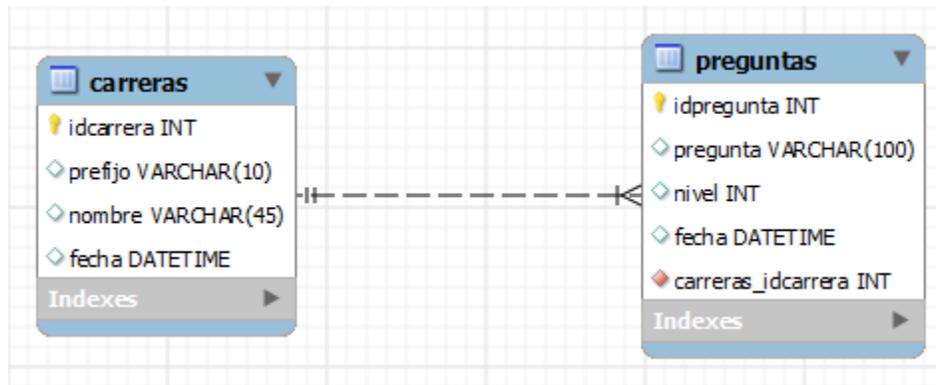


Figura 3. Esquema de la base de datos del sistema

Al terminar de contestar todas las quince preguntas de una pantalla específica, quiz.php envía los valores marcados para dichas preguntas hacia el módulo calculate.php para ser utilizados en el cálculo de las ponderaciones que se requieren y poder determinar el perfil vocacional más apropiado para el usuario. La figura 1 muestra que si la iteración (i) es menor a 4 entonces se regresa el control a quiz.php para desplegar el siguiente bloque de preguntas en una nueva pantalla, no sin antes incrementar el valor de la variable de iteración i para que así todas provengan de un nivel superior. En caso de que la iteración no sea menor a 4 entonces se ejecuta el último módulo results.php para desplegar los resultados de la evaluación, los cuales se explicarán con detalle en la siguiente sección.

Las ponderaciones antes mencionadas son necesarias para determinar el porcentaje de pertenencia de cada pregunta a una carrera específica, es decir, de un total del 100 % para cada pregunta se obtienen porcentajes de pertenencia o asociación por cada carrera. Como se mencionó, la versión inicial del sistema procesa tres carreras, entonces sea $m = 3$ donde m representa el número de carreras que soporta el sistema y n representa el nivel. Por ejemplo, para cada pregunta se asignaron porcentajes correspondientes a las carreras IAS, IEC e ITS, es decir, por medio del cálculo de las ponderaciones es posible determinar si una pregunta pertenece a IAS, IEC o ITS.

La tabla I muestra el procedimiento para calcular las ponderaciones de acuerdo al nivel y el número de carreras soportadas, así como la equidad de las ponderaciones para $n = 0$ debido a que es el nivel de preguntas más generales, las cuales en su mayoría aplican para todas las carreras que procesa el sistema. En cambio, a partir del primer nivel en adelante se debe ver una mayor

ponderación hacia la carrera a la cual pertenece una pregunta. Por ejemplo, para $n = 1$, el valor de la ponderación $1/(3 + 1) = 0.25$ se asigna a cada una de las carreras que no tienen ninguna relación con la pregunta y el valor restante para completar 1 se asigna a la carrera asociada quedando finalmente la distribución $IAS=0.25$, $IEC=0.25$ e $ITS=0.5$ para una pregunta de la carrera de ITS. En general, las ponderaciones mostradas en la tabla I corresponden a preguntas de ITS para cada uno de los diferentes niveles, por ello los valores más altos se encuentran en esa columna. Una gran ventaja es que el sistema se implementó de manera que es capaz de ajustarse a cualquier número de carreras de forma dinámica, lo cual a su vez facilita el cálculo de nuevas ponderaciones sin necesidad de realizar cambios en la lógica del sistema cada vez que se necesite agregar una nueva carrera y su respectivo conjunto de preguntas.

Tabla I. Cálculo de ponderaciones para cada pregunta en base al nivel y al número de carreras

Nivel (n)	Ponderación	IAS	IEC	ITS
0	$1/(m + n)$	0.33	0.33	0.33
1	$1/(m + n)$	0.25	0.25	0.5
2	$1/(m + n)$	0.2	0.2	0.6
3	$1/(m + n)$	0.16	0.16	0.67

Resultados y discusión

Después de presentar una descripción general del sistema de recomendación vocacional *fime.vocacionaluanl*, en esta sección nos enfocamos en la parte final de la operación a lo que son los resultados así como al análisis realizado para validar el desempeño y funcionalidad del sistema. Antes, mostramos en la figura 4 el diagrama de secuencia que sigue el sistema al ser utilizado para poder visualizar los diferentes flujos de operación que normalmente se llevan a cabo entre el usuario, el sistema y la base de datos. Como se puede observar, son tres los principales flujos de operación: mostrar, contestar y ponderar preguntas, los cuales se repiten cuatro veces debido a que son cuatro niveles de preguntas en total (0 al 3). El tiempo promedio que toma responder todo el cuestionario (60 preguntas) es de aproximadamente 10 minutos.

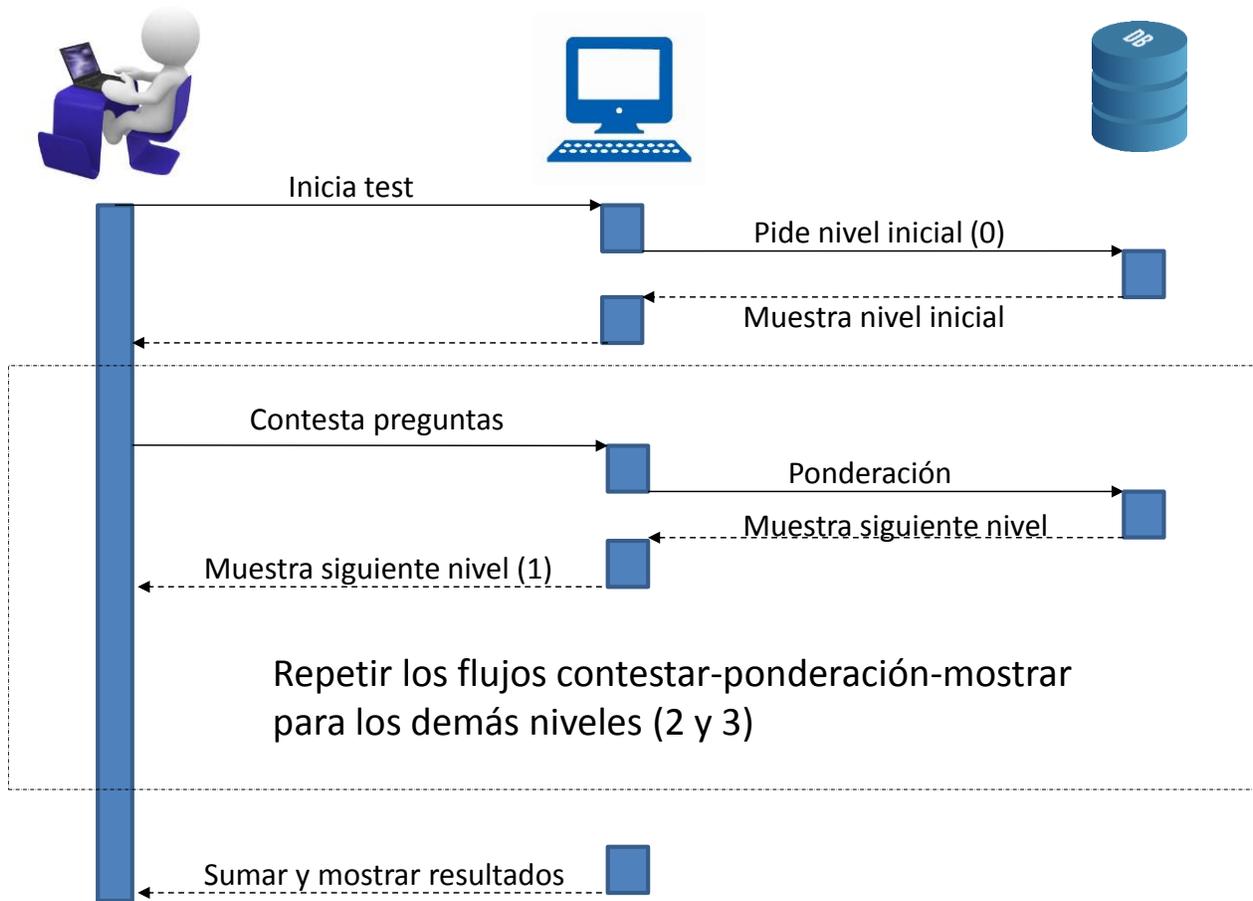


Figura 4. Diagrama de secuencia de operación del sistema *fime.vocacionaluanl*

El sistema se encuentra por el momento disponible en un espacio que se contrató en un servidor para alojar tanto los programas del sistema como su base de datos. Además, el tipo de servicio que se contrató permitiría adaptar el sistema de recomendación a otras facultades de la universidad utilizando el nombre de la facultad seguido del nombre del dominio *vocacionaluanl* para formar su URL. Para el diseño gráfico del sistema, se decidió utilizar Bootstrap el cual es un tipo especial de librería (framework) de software libre diseñado para facilitar la creación de páginas web dinámicas. Basado en los lenguajes HTML y CSS, Bootstrap contiene una gran variedad de funciones que permiten crear aplicaciones web interactivas muy completas.

La figura 5 muestra un ejemplo de cómo aparecen los resultados al terminar de contestar todo el cuestionario. En este caso aparece el nombre de la carrera que el sistema determina más compatible con las habilidades, conocimientos e intereses más desarrollados en el usuario.



Figura 5. Pantalla de resultados del sistema

Por último, es importante terminar presentando un análisis que demuestre de alguna manera que el sistema tiene un desempeño correcto y consistente. Por tanto, dado que el cuestionario representa la parte central de la funcionalidad del sistema de recomendación, para probar la validez y efectividad del mismo se realizó un análisis con el que se obtuvo el índice o coeficiente Alfa de Cronbach para una muestra de usuarios que respondieron el cuestionario. Dicho índice se emplea comúnmente para determinar la consistencia y confiabilidad de cierto instrumento de medición compuesto por varios elementos, variables o ítems correlacionados (Cronbach, 1951). El rango de valores que puede tomar este coeficiente se encuentra entre 0 y 1, donde 0 representa una confiabilidad nula mientras que 1 una confiabilidad absoluta. Por lo regular, un valor igual o mayor a 0.8 se considera una confiabilidad alta.

La muestra analizada ascendió a 40 y fue tomada de una población de estudiantes de las carreras de ITS e IAS principalmente, a quienes se les solicitó usar el sistema para determinar su confiabilidad. Sin embargo, es evidente que el objetivo del sistema es, por un lado, orientar a los alumnos que se encuentran todavía indecisos sobre la carrera que van a estudiar y, por el otro, que el sistema sea capaz de abarcar todas las carreras de ingeniería que ofrece la facultad en versiones posteriores.

Conclusión

En este trabajo hemos presentado un sistema de recomendación vocacional en línea basado en un algoritmo denominado de inteligencia artificial como el principal componente para determinar el mejor perfil vocacional para cada usuario. El sistema tiene como objetivo convertirse en una valiosa herramienta de apoyo de uso libre para todos aquellos estudiantes que estén considerando elegir una carrera de ingeniería en la FIME UANL. El sistema se diseñó e implementó con lenguajes y herramientas de programación ampliamente utilizados en este tipo de aplicaciones dinámicas e interactivas. Adicionalmente se pudo validar la consistencia y fiabilidad del sistema por medio de un análisis estadístico realizado con datos obtenidos de una muestra de 40 usuarios. Se recomienda aumentar en un futuro el número de carreras de FIME que pueda soportar el sistema, e inclusive replicar el sistema de modo que este procese carreras de otras facultades de la UANL. Lo anterior tiene que ver con investigar y definir el conjunto de preguntas que corresponden a todas las carreras que se desea incluir, así como organizarlas en los niveles descritos para que todo se almacene en la base de datos.

Agradecimientos

Deseamos agradecer a los estudiantes de la carrera ITS de la UANL, Xavier Sánchez Díaz, Alfredo González Garza y Cynthia Elizabeth Villagómez Reyna, por su participación en la investigación y desarrollo de la primera versión del sistema de recomendación vocacional; asimismo, agradecemos el apoyo del Programa para el Desarrollo Profesional Docente (PRODEP) de la Secretaría de Educación Pública.

Bibliografía

- Adomavicius, G., y Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp. 734-749.
- Aher, S. B., & Lobo, L. M. R. J. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning system based on historical data. *Knowledge Based Systems*, 51, pp. 1-14.
- Alcaraz, S., Sánchez, X., González, A. (2015). Encuesta orientauanl. FIME UANL. Sitio web: <http://bit.ly/orientaUANL>
- Beati, H. (2011). *PHP: Creación de páginas Web dinámicas*. Buenos Aires: Alfaomega Grupo Editor Argentino.
- Betz, N. E., Fitzgerald, L. F., Hill, R. E. (1989). Trait-factor theories: Traditional cornerstone of career theory. In M. B. Arthur, D. T. Hall, & B. S. Lawrence (Eds.), *Handbook of career theory*. Cambridge: Cambridge University Press, pp. 26-40.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika*, 16(3), pp. 297-334.
- Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. John Wiley & Sons, Inc.
- Franca, D. (2012). Decide tu carrera, orientación vocacional para jóvenes mexicanos. Relpe. Sitio web: <http://www.relpe.org/decide-tu-carrera-orientacion-vocacional-para-jovenes-mexicanos/>
- Gati, I., Fishman-Nadav, Y., Shiloh, S. (2006). The relations between preferences for using abilities, self-estimated abilities, and measured abilities among career counseling clients. *Journal of Vocational Behavior*, 68, pp. 24-38.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities Uand work environments*. Odessa, FL: Psychological Assessment Resources.

- Kasabov, N. K. (1996). Foundations of neural networks, fuzzy systems, and knowledge engineering. Cambridge: MIT Press.
- Katz, M. R. (1993). Computer-assisted career decision making. Hillsdale, NJ: Erlbaum.
- Leavitt, N. (2006). Recommendation Technology: Will It Boost E-Commerce? IEEE Computer Society.
- Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140, pp. 1-55.
- Linden, G., Smith, B., York, J. (2003). Amazon.com Recommendations Item-to-Item Collaborative Filtering, IEEE Internet Computing.
- Sobecki, J. (2014). Comparison of selected swarm intelligence algorithms in student courses recommendation application. International Journal of Software Engineering and Knowledge Engineering, 24(1), pp. 91-109.
- Truemper, K. (2004). Design of logic-based intelligent systems. Wiley-Interscience.
- Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. Second Edition, Morgan Kaufmann.