Artificial Intelligence Techniques for Information Security Risk Assessment

Y. A. Basallo, V. E. Sentí, N. M. Sánchez

***Abstract*— In computer security audits, information security risk (ISR) assessments are performed to computer systems, within it to database management systems (DBMS), often using qualitative methodologies. In these methodologies, the evaluation of the ISR is classified according to its impact in linguistic terms such as: High, Medium or Low, so that ambiguities can be generated in the evaluation result. Security checklists are also used to review the configurations of the DBMS. They have a strong dependence on the presence of the expert auditor in DBMS for this analysis. In order to facilitate the work of the auditors, a model based on knowledge and fuzzy logic was developed for the evaluation of the ISR in the DBMS. In this way, the experience in previous audits of this type is useful and improves the results in the evaluation of the ISR.**

***Keywords*— Audits, artificial intelligence, computer security, risk 6.**

I. INTRODUCCIÓN

Los avances en los sistemas de información (SI) y las tecnologías originan grandes resultados para organizaciones, negocios y otras agencias en términos de productividad del trabajo, almacenamiento de la información, administración y oportunidad de ventajas competitivas. Mientras los SI ofrecen extraordinarios beneficios, también representan mayores niveles de riesgo de modo significativo y sin precedentes, para las operaciones organizacionales. Los negocios, hospitales, escuelas, universidades, agencias gubernamentales y bancos dependen fuertemente de los SI. Esto incrementa la necesidad de la seguridad de la información, según se asegura en [9].

Los gestores de bases de datos son uno de los SI con frecuentes ataques a las vulnerabilidades existentes en las mismas, como exponen en [11]. Se denomina vulnerabilidad a toda debilidad que puede ser aprovechada por una amenaza [5]. La cantidad de vulnerabilidades reportadas al aplicar la inyección de SQL ha ido en aumento en los últimos años según los datos publicados por el Instituto Nacional de Vulnerabilidades de Estados Unidos de América [2]. Este mismo instituto señala que muchos otros tipos de vulnerabilidades de bases de datos se han acrecentado en años recientes. Además se conoce que el 96% de los datos sustraídos durante 2012, provenían de bases de datos, según se publican en [10].

Y. A. Basallo, Universidad de las Ciencias Inofrmáticas, La Habana, Cuba, yazanenator@gmail.com

V. E. Sentí, Universidad de las Ciencias Inofrmáticas, La Habana, Cuba, vivian@uci.cu

N. M. Sánchez, Universidad de las Ciencias Inofrmáticas, La Habana, Cuba, natalia@uci.cu

Corresponding author: Yasser Azán Basallo

Los datos mencionados anteriormente, reflejan la importancia de proteger los datos ante las vulnerabilidades detectadas en los sistemas gestores de bases de datos,

Uno de los pasos para garantizar la seguridad de los datos es la realización de auditorías de seguridad a los sistemas computacionales.

En las auditorías de seguridad a los sistemas computacionales se realiza la evaluación del riesgo de seguridad de la información. La evaluación del riesgo es el proceso de identificación de las amenazas a los sistemas de información, la determinación de la probabilidad de ocurrencia de la amenaza y la identificación de las vulnerabilidades del sistema que podrían ser explotadas por la amenaza [7].

A partir de los estudios realizados, se ha identificado la existencia de dificultades o limitaciones relacionadas con la evaluación del riesgo de seguridad de la información tales como:

- Existen diferentes niveles de experticia entre los auditores para evaluar los sistemas gestores de bases de datos (SGBD) con las listas de chequeo, lo que provoca diferencias entre la evaluación real del riesgo y la estimada por el auditor según encuesta diagnóstico.
- Las listas de chequeo tienen una fuerte dependencia de la opinión del auditor en el resultado del análisis del RSI en los SGBD [8].
- La evaluación del riesgo de seguridad de la informaicón (RSI) en los SGBD se expresa en los términos: Alto, Medio o Bajo, por lo que para cada auditor, constituye una medida ambigua, sin límites precisos.
- Para proporcionar el resultado de la evaluación del RSI, los auditores afirman que se pueden tardar horas o días, por lo que existe demora en la auditoría de seguridad informática y por tanto también en la toma de decisiones.

A partir de los problemas detectados se trazó como objetivo proponer una solución que permita contribuir a mejorar la exactitud en la evaluación del RSI en los SGBD.

II. ESTRUCTURA DEL MODELO PROPUESTO

La palabra modelo proviene del latín modulus que significa medida, ritmo, magnitud y está relacionada con la palabra modus que significa copia, imagen [4].

Se propone un modelo para la evaluación del RSI en los SGBD que pueda utilizar las auditorías pasadas como conocimiento o experiencia plasmada de los expertos que participaron.

Se propone un modelo el cual tiene como entrada, las listas de chequeo de seguridad que emite el Centro para la Seguridad de Internet (CIS) [1], referente a los sistemas gestores de bases

de datos. Documentos reconocidos internacionalmente, que se actualizan cada año y constituyen una guía para los expertos en seguridad informática.

Como salida del modelo, se obtiene la evaluación del RSI en los valores lingüísticos establecidos por los especialistas para un SGBD. Además de esta salida, con el modelo se puede obtener recomendaciones para cada parámetro existente en las listas de chequeo de seguridad del CIS, así como el resultado final de la evaluación de una auditoría, la cual puede contener las evaluaciones del riesgo de varios SGBD.

El modelo se regirá por los siguientes principios:

- La actualización permanente mediante la incorporación de los nuevos casos presentados.
- La flexibilidad para ajustar las variables del riesgo según lo determinen los auditores expertos.
- La estandarización del procedimiento de auditoría a SGBD para la evaluación cualitativa del riesgo.
- La interoperabilidad entre los componentes que conforman el modelo.

Las premisas del modelo propuesto son:

- Disponer de la lista de chequeo de seguridad del CIS para su funcionamiento como entrada.
- Identificar el tipo y la versión del gestor de base de datos con propósito de auditar.
- Los auditores deben revisar los valores cualitativos del riesgo local de los parámetros de la lista de chequeo para corregir alguna imprecisión.

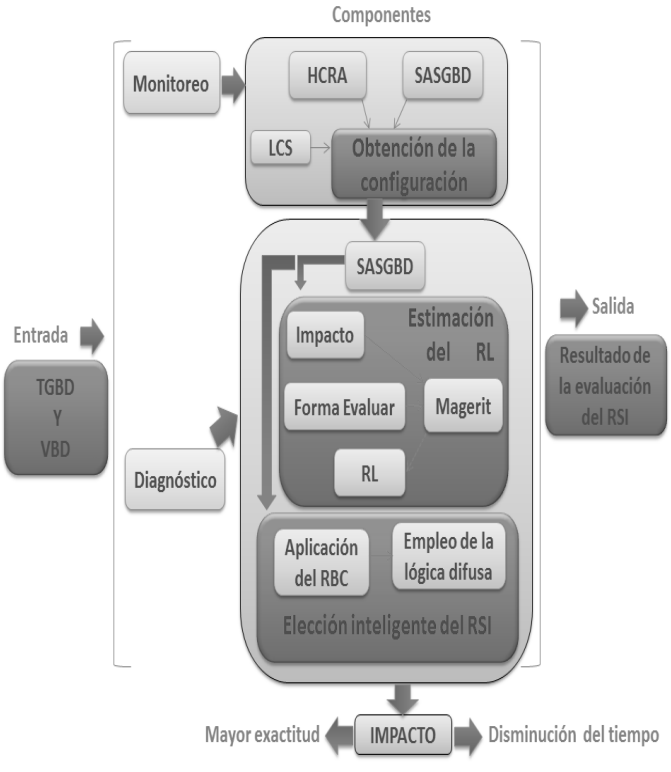

Figura 1. Representación gráfica del modelo.

Componentes del modelo

El modelo está formado por componentes relacionados entre sí como se muestra en la figura 1 y agrupados por las fases: Monitoreo y Diagnóstico.

Los componentes están en diferentes fases porque en diferentes momentos de la auditoría con diferentes requisitos y lugar de trabajo. En la fase Monitoreo, el auditor con la presencia del administrador del SGBD, sustrae las configuraciones de seguridad desde el local sonde se encuentra. Para sustraer las configuraciones de seguridad, se utiliza la lista de chequeo de seguridad del CIS correspondiente al TGBD y a la VBD Se requiere la presencia del administrador del SG BD para que otorgue las credenciales necesarias al auditor para que pueda realizar esta acción y además verifique que durante el periodo de monitoreo, se compruebe que el auditor no realizó una acción que pueda provocar problemas o fallos al SGBD y que los comandos, consultas ejecutadas y alguna otra herramienta puedan ser revisadas por este administrador.

La fase Diagnóstico, es un paso posterior donde no es necesaria la presencia del administrador del SGBD. Sino que se realiza preferencialmente en el local de trabajo del auditor donde puede ser auxiliado por otros auditores y donde son covocadas reuniones de trabajo para analizar la auditoría presente y realizar la toma de decisión con respecto al resultado de la auditoría a presentar.

a) Componente: Obtención de la configuración

Este componente contiene a la Herramienta Colaborativa para la Realización de Auditorías (HCRA), la cual está destinada a apoyar las auditorías que se realizan a través del SASGBD (Sistema de Auditoría para los Sistemas Gestores de Bases de Datos). El HCRA se enfoca en obtener las configuraciones de seguridad del servidor auditado. Las configuraciones están organizadas a través de los parámetros exportados del SASGBD [3] por un archivo con extensión XML. Estos parámetros son los existentes en las listas de chequeo de la CIS.

El HCRA tiene la capacidad de cargar el archivo XML y mostrar las consultas SQL y los comandos a ejecutar que se utilizan para encuestar las configuraciones de seguridad de la base de datos. La solución informática HCRA es capaz de revisar los siguientes SGBD: PostgreSQL, MySQL, SQL Server y Oracle.

El archivo exportado por el HCRA, se convierte en la entrada del siguiente componente del modelo propuesto en esta investigación. Con los datos de configuración introducidos, se crea una matriz de diagnóstico para cada servidor monitoreado a través de la aplicación SASGBD.

b) Componente: Estimación del RL (riesgo local)

Los especialistas entrevistados en el diagnóstico, estiman el RL para cada parámetro de la lista de chequeo de seguridad, solamente con valores lingüísticos: Alto, Medio o Bajo.

En la tabla 1 se muestra como los especialistas realizan el análisis del RL teniendo en cuenta en cuenta las variables: Evaluación del parámetro (evaluación que otorga el especialista según la entrada de HCRA) e Impacto, el valor de esta última proviene de la lista de chequeo.

Tabla 1. Análisis mediante tablas

RL(riesgo local)		Evaluación del parámetro (EP)	
		Bien	**Mal**
Impacto (W)	**Alto**	Bajo	Alto
	Medio	Bajo	Medio
	Bajo	Bajo	Bajo

En la tabla 1 se aprecia el resultado al combinar el impacto y la evaluación del parámetro para determinar el RL, las cuales tienen declaradas las escalas cualitativas como se aprecia en la tabla 2. Está basado en una de las técnicas de la metodología Magerit 3.0 [6].

Tabla 2. Escalas cualitativas seleccionadas de las variables lingüísticas

Impacto (W)	Evaluación del parámetro (EP)
A: Alto	B: Bien
M: Medio	---
B: Bajo	M: Mal

c) Componente: Elección inteligente del RSI

Para la estimación del resultado, este componente se apoya en el uso de las técnicas de la IA: EL razonamiento basado en casos (RBC) y la lógica difusa. El mismo utiliza como entrada lo que corresponde ser la salida del componente anterior, es decir la evaluación del RL de cada parámetro de la lista de chequeo de seguridad empleada y generada como una matriz.

Función de semejanza

Se utiliza una función de semejanza para determinar la semejanza entre casos y de esta forma determinar la evaluación del RSI.

La propuesta general de función de semejanza seleccionada es la publicada en [13]:

Si máx $(|(X_{RN} - X_{RO})|, (|Y_{RN} - Y_{RO}|)) \neq 0$

$$S(RN, RO) = 1 - (1 - \alpha - \beta) * \left(1 - \frac{\int_0^1 \mu_{RN \cap RO}(x)dx}{\int_0^1 \mu_{RN \cup RO}(x)dx}\right) - \alpha \frac{\sum |t_{RNi} - t_{ROi}|}{4} - \beta l_\infty[(X_{RN}, Y_{RN}), (X_{RO}, Y_{RO})] \quad (1)$$

En otro caso:

$$S(RN, RO) = 1 - \left(\frac{1-\alpha-\beta}{2} + \alpha\right) * \frac{\sum |t_{RNi} - t_{ROi}|}{4} - \left(\frac{1-\alpha-\beta}{2} + \beta\right) * |X_{RN} - X_{RO}| \quad (2)$$

La variable RN es el riesgo expresado en un número difuso trapezoidal del nuevo caso de la auditoría de seguridad informática, la cual se desea diagnosticar o evaluar. La variable RO se corresponde al riesgo expresado en un número difuso trapezoidal de un caso almacenado en la BC. La función $S(RN, RO)$ determina la semejanza entre los casos. El valor 1 representa la exacta similitud entre los casos, $\alpha + \beta < 1$, $\mu(R)$ es la función miembro del número difuso R.

$$\beta l_\infty[(X_{RN}, Y_{RN}), (X_{RO}, Y_{RO})] = \underset{\alpha+\beta<1}{máx}((|X_{RN} - X_{RO}|), (|Y_{RN} - Y_{RO}|)), \ \mu_{RN \cap RO}(x) = \underset{[0 \le x \le 1]}{mín}(\mu_{RN}(x), \mu_{RO}(x)), \ \mu_{RN \cup RO}(x) = \underset{[0 \le x \le 1]}{máx}(\mu_{RN}(x), \mu_{RO}(x))) \quad (3)$$

(X_{RN}, Y_{RN}) y $(X_{RO}, Y_{RO}$ Son los centroides de RN y RO y se calculan de la siguiente manera según se publica en [12]:

$$X_R = \left\{Y_{R(t_3+t2)+(w_R - Y_R)(t_4+t_1)}\right\},$$

$$Y_R = \left\{ \begin{array}{l} \frac{\left(\frac{t_3-t2}{t_4-t_1}\right)}{6}, \text{si } t_4 - t_1 \neq 0 \\ \frac{1}{2}, \text{ si } t_4 - t_1 = 0 \end{array} \right\} \quad (4)$$

Las variables α y $\beta = {}^1/_3$ para que se puedan comparar los resultados con el análisis dispuesto en [12].

$$R = (\sum_{i=1}^{n} peso_i \otimes RL_i) ø (\sum_{i=1}^{n} peso_i) \quad (5)$$

En la anterior ecuación (5), la variable R que representa el riesgo del servidor con un número difuso trapezoidal generalizado R= (t1, t2, t3, t4; w). Son números reales t1, t2, t3, t4 y w tal que: $0 \le t1 \le t2 \le t3 \le t4 \le 1$, $0 \le w \le 1$. La variable w representa la altura del trapecio.

Para aplicar la ecuación (5) hay que convertir los valores cualitativos de RL a números difusos, La asociación quedó de la siguiente manera representado en la tabla 3.

Tabla 3. Representación difusa de los valores lingüísticos

Valores lingüísticos	Números difusos trapezoidales
Alto	(0.6, 0.66, 1, 1; 1)
Medio	(0.31, 0.37, 0.59, 0.65; 1)
Bajo	(0, 0, 0.3, 0.36; 1)

Ejemplo de determinación del número difuso R:

Para determinar el número difuso R a través de la ecuación (5) se necesita la entrada del componente Estimación del RL, de una lista de chequeo para el SGBD PostgreSQL, el cual puede quedar según la tabla 4.

Tabla 4. Fragmento de evaluación del RL de parámetros de una lista de chequeo de un SGBD PostgreSQL (Los pesos son solo ejemplos)

NOMBRE DEL PARÁMETRO	RL	Peso
Actualización del catálogo del sistema	Alto	0.69
Pertenencia de usuarios a grupos.	Medio	0.47
Usuarios con claves nulas.	Alto	1
Cuentas vencidas	Bajo	0.3

Para el ejemplo de la tabla 4, la ecuación (5) queda de la siguiente forma:

$$R = [(0.6, 0.66, 1, 1; 1) \otimes 0.69) \oplus (0.31, 0.37, 0.59, 0.65; 1) \otimes 0.47 \oplus (0.6, 0.66, 1, 1; 1) \otimes 1 \oplus (0, 0, 0.3, 0.36; 1) \otimes 0.3] \text{ø}(0.69 + 0.47 + 1 + 0.3)$$
$$= [(0.414; 0.4554; 0.69; 0.69; 1) \oplus (0.1457, 0.1739, 0.2773, 0.3055; 1) \oplus (0.6, 0.66, 1, 1; 1) \oplus (0, 0, 0.09, 0.108; 1)] \text{ø}(2.46)$$
$$= [(1.1597, 1.2893, 2.0573, 2, 1035; 1)] \text{ø}(2.46)$$
$$= (0.47, 0.52, 0.84, 0.86; 1)$$

III. RESULTADOS

Se aplicó un experimento para probar la contribución del modelo propuesto a través de las aplicaciones informáticas SASGBD y HCRA en la investigación a partir del estudio de casos.

La aplicación del experimento tiene como objetivo comparar los resultados de la evaluación del RSI obtenidos por la instancia del modelo a través de la lógica difusa y el RBC con los casos de estudio (Ox) proporcionados por especialistas, donde 0<x<16.

En la columna valor esperado de la tabla 5, es el valor del resultado de la evaluación de RSI de los casos de estudio proporcionados por los especialistas. La columna valor observado, es el valor proporcionado por la solución informática compuesta por el HCRA y el SASGBD, las cuales son la instancia del modelo. La etiqueta que tenga el valor de semejanza más alto, será la etiqueta de la respuesta del modelo.

Se toma como valor **correcto del modelo**, donde coincida el valor observado con el valor esperado.

Los aceptables son casos aquellos donde no coinciden estsa columnas, pero el valor del observado es un resultado admisible por los especialistas que entegaron los casos de estudio. Los valorados de Mal son aquellos que los valores no coinciden y no son aceptados.

Los resultados de la medición de la exactitud del experimento, mostrados en la tabla 5 y figura 2 son que: de los 15 casos, 10 fueron evaluados correctamente, ninguno sin evaluar, 3 evaluados incorrectamente y 2 de modo aceptable. Obteniendo como correctos un 67% de los casos de estudios. Si se toman los casos evaluados correctamente y los aceptables como casos bien evaluados, entonces el modelo es capaz de acertar en un 87% para los casos de estudios, elevando su exactitud.

Tabla 5. Resultado del experimento.

Caso	Semejanza Alto	Semejanza Medio	Semejanza Bajo	Valor observado	Valor esperado
O_1	0.620	0.500	**0.830**	Bajo	Bajo
O_2	0.620	0.500	**0.830**	Bajo	Alto
O_3	0.833	**0.957**	0.400	Medio	Medio
O_4	0.786	**0.900**	0.383	Medio	Alto
O_5	0.830	**0.833**	0.510	Medio	Alto
O_6	**0.830**	0.687	0.619	Alto	Alto
O_7	**0.800**	0.660	0.643	Alto	Alto
O_8	0.721	**0.782**	0.543	Medio	Alto
O_9	0.398	0.429	**0.940**	Bajo	Bajo
O_{10}	0.334	0.367	**0.863**	Bajo	Bajo
O_{11}	0.370	0.400	**0.940**	Bajo	Bajo
O_{12}	0.740	**0.810**	0.560	Medio	Alto
O_{13}	**0.935**	0.849	0.377	Alto	Alto
O_{14}	0.557	0.659	**0.716**	Bajo	Alto
O_{15}	**0.849**	0.837	0.462	Alto	Alto

Los casos de estudio valorados como correctos son aquellos donde coincide el valor esperado con el observado. Los aceptables son casos aquellos donde no coinciden los valores, pero el valor del observado es un resultado admisible. Los valorados de Mal son aquellos que los valores no coinciden y no son aceptados.

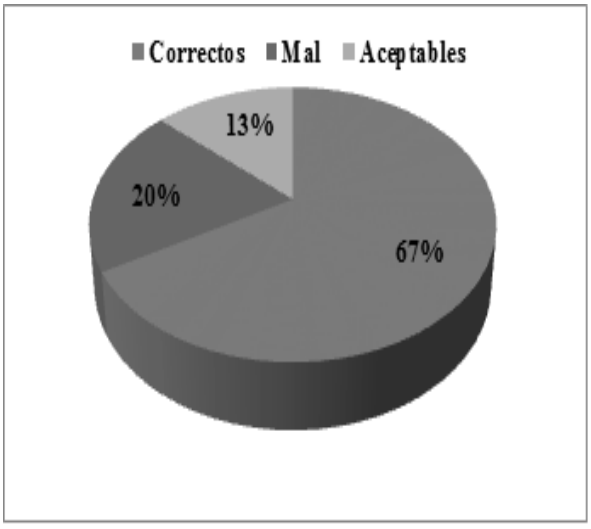

Figura 2: Resultados de la medición de la exactitud del modelo.

IV. CONCLUSIONES

La investigación realizada permite llegar a las siguientes conclusiones:

- Con el experimento realizado se evidencia la viabilidad de las técnicas de inteligencia artificial: RBC y lógica difusa para realizar la evaluación del RSI en los SGBD.
- Se logra evaluar los diferentes casos de estudio diseñados para el experimento, por lo cual se refleja la capacidad del modelo para responder a diferentes auditorías con disímiles características.
- Los resultados del experimento muestran la posibilidad del modelo para contribuir en la mejora de la exactitud en la evaluación del RSI en las auditorías de seguridad informática a los SGBD hasta de un 87% si se cuentan como bien, los casos de estudios evaluados de correctos y aceptables. Por lo que para los auditores en seguridad informática de los SGBD, repercute en una mejor exactitud en la estimación del RSI.

V. REFERENCIAS

[1]. "C.I.S.". "CIS Benchmarks". pp. Number of Pages. 2013. Fecha de consulta:12/11/2013. http://benchmarks.cisecurity.org/downloads/multiform/.

[2]. "N.I.S.T.". "Datos estadísiticos de Inyección de SQL ". pp. Number of Pages. 2017. Fecha de consulta:1/1/2017. http://web.nvd.nist.gov/view/vuln/statistics-results?cves=on&query=&cwe_id=CWE-89&pub_date_start_month=-1&pub_date_start_year=2007&pub_date_end_month=-1&pub_date_end_year=-1&mod_date_start_month=-1&mod_date_start_year=-1&mod_date_end_month=-1&mod_date_end_year=-1&cvss_sev_base=&cvss_av=&cvss_ac=&cvss_au=&cvss_c=&cvss_i=&cvss_a=.

[3]. Y. Azán Basallo y otros. "Solución basada en el Razonamiento Basado en Casos para el apoyo a las auditorías informáticas a bases de datos".Revista Cubana de Ciencias Informáticas. Edition. City. Vol. 8, pp. Number of 52-68. 2014.DOI:

[4]. A.L. Del Valle. "Metamodelos de la investigación pedagógica". La Habana (Cuba). VolI. pp. Number of Screens. 2007. ISBN:

[5]. M.H.A.P. "MAGERIT – versión 3.0 Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información". MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS. Madrid (España). VolI. 1, pp. 42. 2012.

[6]. M.H.A.P. "MAGERIT – versión 3.0 Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información". MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS. Madrid (España). VolI. 3, pp. 42. 2012.

[7]. M.-S.I.S.a.A.C. MS-ISAC. " Cyber Security: Risk Management A Non-Technical Guide Essential for Business Managers Office Managers Operations Managers ". pp. Number of Pages. 2010. Fecha de consulta:14/12/2013. http://msisac.cisecurity.org/resources/guides/documents/Risk-Management-Guide.pdf#page=3&zoom=auto,179,0.

[8]. M.G.V. Piattini y E.N. De Peso. "Auditoría Informática. Un enfoque práctico". RA-MA Editorial. 2da. Madrid (España). VolI. 1, pp. 2001. ISBN:H4-7897-444-X.

[9]. M. Quigley. "Encyclopedia of Information Ethics and Security". Information Science Reference. illustrated. VolI. pp. 2008. ISBN:9781591409885.

[10]. A.E.V. Quisbert. "Modelo de Sistemas Multi-Agentes para Percibir, Evaluar y Alertar Ex-Antes los Accesos no Autorizados a Repositorios de Base de Datos". Revista del Postgrado en Informática. pp.128.2014. ISSN:3333-7777

[11]. D. Ramakanth y K. Vinod. "SQL Injection - Database Attack Revolution And Prevention". Journal of International Commercial Law and Technology. **6.** pp.224-231.2011. ISSN:19018401

[12]. E. Vicente, A. Mateos, y A. Jiménez. "A new similarity function for generalized trapezoidal fuzzy numbers". in *International Conference on Artificial Intelligence and Soft Computing*. 2013: Springer.

[13]. E.C. Vicente, A.C. Mateos, y A.M. Jiménez "Un enfoque borroso para el análisis y la gestión de riesgos en sistemas de información".Journal. pp. 106. 2013.

Yasser Azán Basallo es ingeniero en ciencias informáticas, egresado de la Universidad de las Ciencias Informáticas en La Habana, Cuba, desde el 2007. Recibió el máster en diciembre del 2012 en el mismo centro universitario en el cual se desempaña como profesor. Su área de investigación está relacionada con la seguridad informática y la inteligencia artificial.

Natalia Martínez Sánchez es doctora en Ciencias Técnicas, desde el 2009 en la Universidad Central "Marta Abreu de las Villas". Profesora Titular de la Universidad y Vicerrectora de Formación de la Universidad de las Ciencias Informáticas de Cuba.

Vivian Estrada Senti es doctora en Ciencias Técnicas, Especialidad. Profesora Titular de la Universidad de Ciencias Informáticas de Cuba. Ha formado Doctores en Ciencia en las áreas de Informática y Ciencias de la educación. Pertenece al Comité Internacional Coordinador de la Cátedra UNESCO. Ha impartido conferencias y cursos en varios países como España, México, Brasil, R. Dominicana, Guatemala, Colombia y en eventos de la Unión Europea. Es miembro del Tribunal Nacional de Grados Científicos de Ciencias de la Educación.