Método gráfico para la validación de identidad humana en usuarios de la web.

Graphic method for human validation to identity web users.

José Alberto Noh Noh (1).
Estudiante, Universidad Autónoma de Yucatán, Facultad de Matemáticas.
jose31994@gmail.com.

Cinhtia Maribel González Segura (2). Universidad Autónoma de Yucatán, Facultad de Matemáticas.
gsegura@correo.uady.mx.

Artículo recibido en octubre 14, 2015; aceptado en noviembre 10, 2015.

Resumen.

Los métodos de autentificación denominados CAPTCHA se han vuelto las pruebas humanas iterativas más utilizadas para la autentificación de usuarios en la web, ya que permiten la diferenciación entre usuarios humanos y algoritmos automatizados, con el fin de brindar seguridad a la información de los sitios web que lo requieren. En este trabajo se presentan algunos conceptos relacionados con las pruebas humanas iterativas, posteriormente se describe el desarrollo de un nuevo método CAPTCHA propuesto, basado en la identificación de imágenes gráficas seccionadas, el cual es comparado con un método tradicional basado en el clásico reconocimiento de caracteres. Se realiza también un análisis comparativo de ambos métodos, el cual consiste en que un grupo de usuarios evaluaran la usabilidad del sistema después de haber interactuado con el sistema desarrollado que incorpora un método basado en OCR y otro NO basado en OCR. Los resultados muestran que el método propuesto resulta de mayor agrado para los usuarios, a la vez que garantiza el acceso seguro para los usuarios humanos de la web.

Palabras clave: CAPTCHA, seguridad de información, clasificación fundamental, pruebas humanas iterativas, método propuesto, OCR, NO OCR, usabilidad.

Abstract.

The CAPTCHA authentication methods have become the iterative human tests most used for user authentication on the Web. They allow differentiation between human users and automated algorithms, in order to provide security information websites that is required. In this work, some iterative human testing concepts are presented; then, the developing of new CAPTCHA proposed method is described, based on the identification of graphic images sectioned, which is compared to a traditional method based on the classical character recognition described. A comparative analysis of the two methods, which is a group of users assess the usability of the system after having interacted with the developed system incorporating a system based on OCR and other NO-based OCR method is also performed. The results show that the proposed method is most pleasing to the users, while ensuring secure access to human users of the web.

Keywords: CAPTCHA, information security, basic classification, human iterative proofs, method proposed, OCR, NO OCR, usability.

1. Introducción.

Los sitios web que ofrecen formularios de registro para acceder a servicios gratuitos o de pago, suelen sufrir ataques con software dañino que algunos usuarios de la red colocan en dichos sitios, el acceso común de este software ocurre

cuando se realiza una solicitud al servidor que aloja al sitio web, por lo que al ingresar al servidor causa daños o problemas que van desde la eliminación de páginas hasta el robo de identidad o la denegación del servicio (saturación del servidor), este último debido a la realización automática de la misma solicitud al servidor varias veces. Esto da origen a los métodos de autenticación, un pequeño candado de fácil solución para un ser humano pero difícil para una computadora, logrando así regular el acceso a ciertos servicios de la web.

Tales métodos provienen de técnicas que se han denominado pruebas humanas iterativas, mejor conocidas como HIP (Huma Interactive Proofs), permiten a una persona autenticarse, mediante un desafío que el computador ofrece, el cual debe ser fácil de superar por un humano, pero difícil para quien no lo sea (Areitio, 2007), en otras palabras, el usuario humano demuestra serlo través de un desafío/respuesta de protocolo (Shirali-Shahreza, 2008), quienes, además, señalan que la mayor parte de los HIP son de tipo grafico(palabras, imágenes e inclusive videos).

Desde sus inicios, los métodos de autenticación de usuarios costaban tiempo y esfuerzo, convirtiéndose en una tarea complicada y tediosa para los usuarios. Así, diseñadores y programadores web se dieron a la tarea de mejorar los métodos de autenticación de usuarios, llegando a lo que en la actualidad se conoce como CAPTCHA, el cual es uno de los métodos de autenticación más populares en la actualidad.

El auge de los sitios web y sistemas en línea ha hecho necesaria la creación de soluciones de seguridad que permitan restringir el acceso a programas maliciosos o bots mediante la autenticación de la identidad humana. Usando la autentificación denominada CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*), los cuales están basados en inteligencia artificial (Shirali-Shahreza, 2008), Los CAPTCHA son denominados como un "*puzzle*" que los webmasters incluyen en su sitio web para asegurarse de que los visitantes que quieren interactuar con el contenido son personas, y no robots spam (Cabezas, 2014).

Los CAPTCHA, parten de la inteligencia artificial, por lo que "benefician al propietario de un sitio web porque filtran los no deseables robots spammers, y de paso pueden proporcionar mayor percepción de seguridad al usuario" (Cabezas, 2014), y proponen retos basado en texto o imágenes, aptos para ser rápidamente resueltos por seres humanos pero difíciles para las computadoras, la aplicación de uno de los métodos CAPTCHA más populares en la actualidad consiste en presentar diversos caracteres distorsionados en una imagen cuyos efectos pictóricos dificultan ligeramente la identificación del código que el usuario humano debe digitar y que no puede ser reconocido fácilmente por un bot malicioso.

Sin embargo, el término CAPCHA se comienza a utilizar en el año 2000 en la Universidad de *Carnegie Mellon*, y responde a un juego de palabras, ya que la pronunciación de la palabra recuerda a *catch ya*, una versión informal de *I catch you* (Martinez, 2009).

Los CAPTCHA, se clasifican fundamentalmente en métodos basados y no basados en OCR (*Optical Character Recognition*) (Shirali-Shahreza, 2008), los métodos basados en OCR, son los más conocidos y usados en la actualidad, se presenta la imagen de una palabra con una distorsión de diversos efectos, la cual debe ser escrita por el usuario y, debido a los efectos pictóricos, no podrá ser reconocida por el equipo. Para su creación, se acostumbra llevar a cabo un procedimiento general, en el que se diferencian los métodos de generación por medio de los algoritmos de elección de palabras/diccionarios, el formato aplicado a los caracteres y las degradaciones realizadas sobre las imágenes.

Sin embargo, el método común basado en OCR presenta algunos inconvenientes tales como no lograr comprender el texto de la imagen, colores sin contraste que dificultan su lectura, vulnerabilidad si la distorsión no es correcta, mayor dificultad para personas sin experiencia en computación etc.

Los métodos no basados en OCR presentan imágenes con retos cuya resolución implica actividades tales como dar clic en ciertas zonas específicas de la imagen, identificar una serie oculta en las imágenes, mover algún componente de la imagen, o incluso formar cadenas de caracteres con las iniciales de los objetos representados. Una de las principales razones por las que los CAPCHA basados en imágenes son vulnerables frente a ataques, es que en caso ninguna de las técnicas existentes esta es distorsionada para evitar el reconocimiento de una máquina.

En el método no OCR que se explicará más adelante, un usuario debe identificar las partes incongruentes de una imagen, en la que cuyo algoritmo particiona la imagen y coloca 2 piezas aleatorias que solo un usuario humano podría resolver, de esta manera el usuario podrá validar su identidad humana restringiendo así el acceso automatizado de bots a los sistemas web. Se realizó una comparación del método propuesto y un método OCR, con el propósito de comparar la usabilidad y opinión de los usuarios.

2. Métodos.

De acuerdo con Cabezas, Sabate, Vendrell y Marcos (2014) los CAPTCHA son unos rompecabezas o retos que los administradores de sitios web incluyen para asegurarse de que los visitantes que interactúan con el contenido son personas y no robots automatizados que tratan de acceder al sitio web para obtener o añadir información de manera indiscriminada, dañando la integridad del sitio web.

En sus inicios, los primeros métodos CAPTCHA fueron basado en OCR, Udi Manber (Jefe científico, 1998) de Yahoo! Presenta a los BOTS (programa informático que realiza distintos contenidos y que trata de simular a un humano) y la necesidad de evitarlos en los chat, para lo que los profesores Manuel Blum, Luis A. von Ahn y John Langford desarrollan un *gimpy*, con palabras en ingles, al azar, presentando como una imagen de texto impreso con un amplia variedad de deformaciones y distorsiones, incluyendo las imágenes superpuestas de palabras diferentes (Hernández, 2010), Sin embargo, el método común basado en OCR presenta algunos inconvenientes tales como no lograr comprender el texto de la imagen, colores sin contraste que dificultan su lectura, vulnerabilidad si la distorsión no es correcta, mayor dificultad para personas sin experiencia en computación etc.

Figura 1. Imagen generada con el software Gimpy al elegir siete palabras de un diccionario y crear una imagen distorsionada que contiene las palabras.

Durante los inicios de los métodos no basados en OCR (gráficos), uno de los métodos implementados fue denominado *Bongo*, En este método se le realiza preguntas al usuario, para resolver un problema de reconocimiento de patrones visuales, en la que se mostraban dos series de bloques, en la que la serie de la derecha dependía de la de la izquierda y viceversa, en este tipo de métodos, un usuario debe identificar las partes incongruentes de una imagen, lo cual permitirá validar su identidad humana restringiendo así el acceso automatizado de bots a los sistemas web.

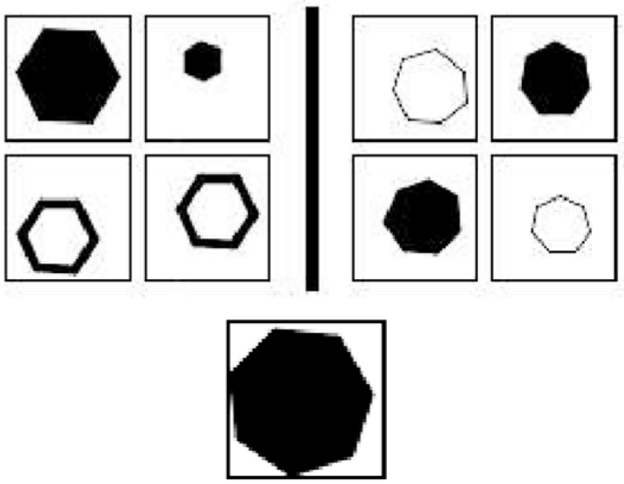

Figura 2. Ejemplo de ejercicio generado con Bongo, un método grafico donde el usuario debe elegir la pieza que no corresponde a los patrones de ambas series presentadas.

Considerando lo anterior, se diseñó e implementó un método no basado en OCR, para la autentificación de usuarios, cuyo propósito es el de disminuir el tiempo de acceso, mejorar la experiencia de los usuarios y crear un método fácil de resolver por un usuario humano, y difícil para una computadora.

Se realizó una comparación de un método propuesto No basado en OCR, usando lenguajes de desarrollo web y un método común de tipo textual que es basado en OCR, con el propósito de comparar la usabilidad y eficiencia, este método a comparar fue obtenido del sitio http://www.phpcaptcha.org, y adaptado en un sitio para la realización de pruebas.

3. Desarrollo.

Se desarrolló un sistema web en el que se implementaron dos métodos de autentificación: uno basado en OCR y otro no basado en OCR.

En el primer método, un algoritmo muestra de manera aleatoria una imagen seccionada en una matriz que simula ser un rompecabezas en el que existen dos piezas mal posicionadas, las cuales el usuario debe reconocer y al hacer clic sobre las piezas correctas se le concede el acceso. Una variante del método podría incrementar la seguridad añadiendo más de dos piezas. Inicialmente se utilizaron 30 imágenes elegidas de manera aleatoria, con características óptimas para ser resueltas por los usuarios, las cuales se almacenaron en un servidor en línea para las pruebas, al igual que los algoritmos y/o códigos implementados.

De manera general se usaron códigos del lenguaje Canvas de JavaScript, para la división (Algoritmo 1, se usaron 2 arreglos de forma global arreglopiezas y arreglocorrecto, estos arreglos serán importantes y serán usados por otras funciones), intercambio de piezas(Algoritmo 2, es necesario arreglopiezas, en este algoritmo se lleva a cabo el intercambio de las dos piezas necesarias para nuestro método propuesto) y verificación (Algoritmo 3, se usaran los arreglos arreglopiezas y arreglocorrecto, para realizar una comparación entre estos), además de códigos propios del lenguaje para la ubicación del orden correcto de cada pieza y la colocación y división de la imagen en trozos del mismo tamaño, según sea la cantidad de divisiones por lado.

Algoritmo 1. Algoritmo de división

```
Division(tamañoimagen, divisiones, imagen)
{
    tamañobloque = tamañoimagen/divisions;
    var r;
    for (var i=0; i < divisions; i++)
    {
        for(var j=0; j< divisions;j++)
        {
            r = new Rectangle(i * tamañobloque, j * tamañobloque, i* tamañobloque + tamañobloque, j *
tamañobloque + tamañobloque);
            arreglopiezas.push(r);
            arreglocorrecto.push(r);
        }
    }
}
```

Algoritmo 2. Algoritmo de intercambio

```
Intercambio(arreglopiezas)
{
        var count = 0;
        var temp;
        var index1;
        var index2;
        while(count < 1)
            {
                index1 = Math.floor(Math.random()*arreglopiezas.length);
                    do{
                        index2 = Math.floor(Math.random()*arreglopiezas.length);
                    }while(index1===index2);
                temp = arreglopiezas [index1];
                arreglopiezas [index1] = arreglopiezas [index2];
                arreglopiezas [index2] = temp;
                count++;
            }
}
```

Algoritmo 3. Algoritmo verificacion

```
verificacion()
{
        var match = true;
        for(var i = 0; i < arreglopiezas.length; i++)
            {
                if(arreglopiezas [i] != arreglocorrecto[i])
                    {
                        match = false;
```

```
                                        }
                }

        if(match)
                {
                        console.log('Finalizado');
                }
                else
                {
                        console.log(No finalizado);
                }

}
```

Un ejemplo seria la división de una imagen en cinco particiones por lado, generaría, una imagen con 25 piezas del mismo tamaño, cada pieza es almacenada en un arreglo, usando una función aleatoria, se escoge dos números distintos según el tamaño del arreglo, y son intercambiados, almacenándose en un nuevo arreglo con el nuevo orden, seguidamente se vuelve a mostrar la imagen, pero ahora correspondiente al nuevo arreglo, mostrándose así las dos piezas intercambiadas, de manera específica, existe una solución, y esta es dar clic en las dos piezas mal colocadas, por el momento permite seleccionar otras pero no serán una solución esto llevaría al usuario demorarse en solucionarse, a futuro, para evitar este problema estaría como margen el tiempo en la que se es posible resolver el método y cambiar o actualizar a imagen a una nueva con otro orden, en este método se evalúa el tiempo en la que el usuario dura en solucionarlo.

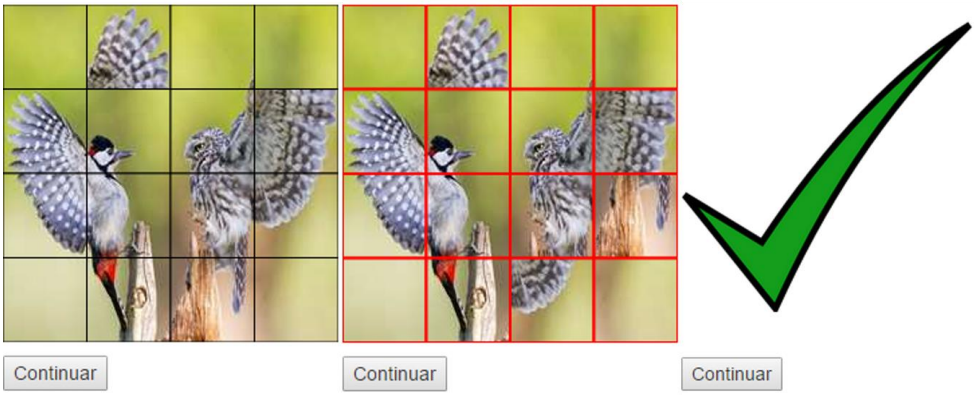

Figura 3. Método No OCR Propuesto, Inicialmente contará con dos piezas mal colocadas, si comete errores, cada pieza es rodeada por líneas rojas, lo cual es incorrecto, y le tomara tiempo al usuario, si es resuelta de manera correcta aparecerá la tercera imagen.

El segundo método basado en OCR implementado consiste en una interfaz tradicional de caracteres alfanuméricos distorsionados que el usuario debe reconocer y transcribir idénticamente para validar su identidad humana, este fue obtenido de un sitio web, el cual desarrolla varias variantes de este captcha captcha, con la comprensión visual y la comprensión por medio de audio, se tomó la forma básica del método obtenido, el cual fue adaptado a una página web con formulario y la opción de cambiar la imagen, en este método se evalúa el tiempo, en la que el usuario dura en resolver una imagen, si cambia la imagen, el tiempo inicia nuevamente, este método fue de igual manera adaptado al sitio de pruebas y cargado al servidor en línea para las pruebas.

Figura 4. Método basado en OCR obtenido, muestra una imagen con caracteres alfanuméricos que deben ser ingresados en el campo textual para poder realizar la verificación del método, en este método es posible cambiar la imagen que se muestra.

Ambos se diseñaron utilizando HTML, PHP y JavaScript, así como un servidor de páginas web Apache y un servidor de Hosting en línea. La interfaz completa consiste en una secuencia de formularios que permiten al usuario interactuar con ambos métodos consecutivamente y expresar su opinión al respecto de su experiencia. Así, el novedoso método propuesto podrá ser incluido en cualquier sitio web en que desee asegurarse de que los visitantes que interactúan con él sean personas y no robots spam que tratan de registrarse en el sitio web, incluir comentarios en blogs, etc. (Cabezas, 2014).

De acuerdo con (Alicia Yesenia Lopez Sanchez, 2013), para un ser humano, el tiempo promedio de resolución de un método de autentificación basado en OCR es de 16 segundos, y para un bot el tiempo promedio que le toma obtener un intento de solución, que puede ser errónea o correcta, es de aproximadamente 6 segundos, independientemente del método empleado (Saputra C, 2013).

Una vez que el sistema se publicó en un servidor web en línea, la experimentación se realizó difundiendo la dirección e invitando a participar en diversos medios digitales. Se contó con la participación de 21 personas, quienes interactuaron con ambos métodos y posteriormente emitieron su opinión al respecto, de los cuales el 67% manifestó poseer una experiencia intermedia y un 33% dijo contar con un nivel avanzado en el uso de sistemas.

Se omite a una persona para el resultado promedio del tiempo de resolución de los métodos cuya información no es confiable, de las 10 personas restantes se obtiene que en conjunto los tiempos de resolución para el método propuesto No basado en OCR fue en promedio de 6 segundos, mientras que el segundo método basado en OCR, en promedio fue resuelto en aproximadamente 12 segundos, los usuarios de igual manera opinaron que el primer método (OCR), era más de su agrado con un 100% de aceptación a comparación del segundo (NO OCR), que para algunos era complicada su resolución. El sistema desarrollado se encuentra disponible en la dirección: http://www.pruebacaptcha.esy.es.

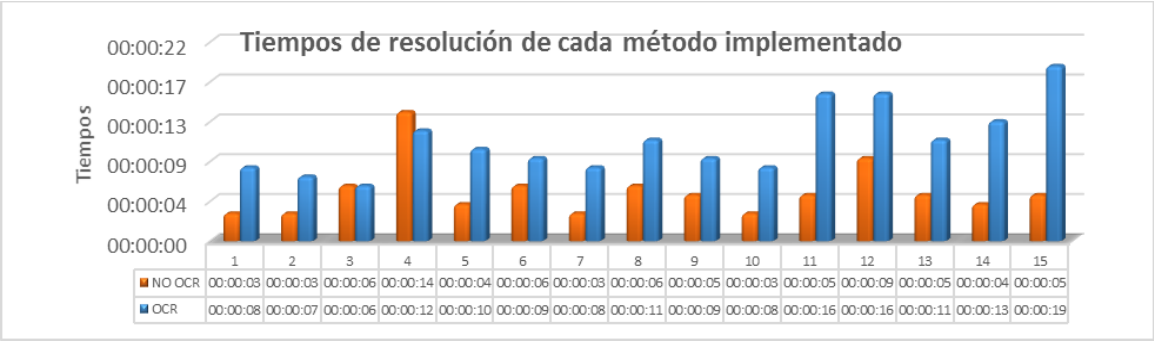

Figura 5. Tiempo de resolución de ambos métodos por usuario, Se puede notar de manera clara que el método propuesto es resulto en menor tiempo.

En la figura 5 se puede observar de manera clara que el método No basado en OCR que se ha propuesto fue resulto en un tiempo menor a comparación del método basado en OCR, en la gráfica antes mencionada se muestra la distribución de tiempos de resolución, de 10 pruebas realizadas, en donde se puede notar la diferencia en tiempos de resolución de cada método que se implementó en la prueba hecha.

Conclusiones.

Con la información anterior se concluye que el método basado en OCR propuesto requiere menor tiempo de solución para un humano y para los usuarios resulta ser de mayor agrado que el método tradicional textual que comúnmente se encuentra en diversos sitios web.

Posteriormente, se sugiere realizar evaluaciones de usabilidad para garantizar que las personas puedan resolverlas en tiempo y tazas de éxito razonable (Gossweiler, 2009).

Con la información recabada es importante mencionar que el método que se ha desarrollado para realizar la comparación ha sido, aceptado por el 100% de los usuario, además de que estos mismo han comentado que el método No OCR, puede ser resuelto en un tiempo mucho menor, y ofrecer la seguridad necesaria a la de un método común, este método es más atractivo para los usuario, e inclusive divertido de resolver, a varios usuario han demostrado el gusto de la resolución por la imágenes que estas contienen.

De esta manera se concluye que el método basado en No OCR desarrollado y propuesto, puede ser implementado como una buena alternativa al método basado en OCR con el cual se ha realizado comparación.

Referencias Bibliográficas.

Alicia Yesenia Lopez Sanchez, N. U. (2013). Comparación De Usabilidad Entre Captcha Basado En Texto Y Captcha Basado En Imágenes. *Memorias Arbitradas Del VIII Congreso De Ingeniería Industrial Y De Sistemas* (págs. 48-58). San Nicolas de los Garza, Nuevo Leon, Mexico: Facultad De Ingeniería Mecánica y Eléctrica.

Areitio, J. y. (2007). Análisis en torno a la tecnología biométrica para los sistemas electrónico de identificación y autentificación. *Revista española de electronica(630)*, 52-67.

Cabezas, V. S. (2014). Experiencia de usuario y captchas, explorando la semiotica visual. no solo usabilidad: *revista sobre personas, diseño y tecnologia*, 1-14.

Gossweiler, R. K. (2009). Whats Up CAPTCHA A CAPTCHA Based On Image Orientation. *18th international conference on World wide web (págs. 841-850). Nueva York, NY, EE.UU: 18th international conference on World wide web*.

Hernández, C. J. (2010). Pitfalls in CAPTCHA design and implementation: The Math CAPTCHA, a case study. *Computers & Security*, 141-157.

Martinez, D. y. (2009). Servicios Accesibles de Acceso Esclusivamente Humano. *Segunda Conferencia Internacional sobre brecha digital e inclusión social 9*, (págs. 1-12). Madrid, España.

Saputra C, A. E. (2013). Comparison of Classification Algorithms to tell Bots and Humans Apart. *Journal Of Next Generation Information Technology [serial online]*, 23-32. Obtenido de http://search.ebscohost.com/login.aspx?direct=true&db=aci&AN=98906315&lang=es&site=ehost-live

Shirali-Shahreza, M. y.-S. (2008). Encouraging persons with hearing problem to learn sign language by Internet websites. *Eighth IEEE International Conference on Advanced Learning Technologies, ICALT '08 , IEEE* (págs. 1-3). Tehran, IRAN: Eighth IEEE International Conference on Advanced Learning Technologies.

Información de los autores.

Jose Alberto Noh Noh, es estudiante de la Licenciatura en Ciencias de la Computación de la Facultad de Matemáticas en la Universidad Autónoma de Yucatán, sus principales intereses son las áreas de programación de sistemas, desarrollo web y móvil. ha participado en el verano de investigación científica de la península de Yucatán (2015) Priiori-CONACYT, actualmente imparte cursos en las áreas de matemáticas y mantenimiento de equipos de cómputo. El artículo presentado es parte de su trabajo de tesis.

Cinhtia Maribel González Segura es Maestra en Ciencias de la Computación por el Instituto Tecnológico y de Estudios Superiores de Monterrey (2005). En agosto de 2014 inició el Doctorado en Sistemas y Ambientes Educativos en el Sistema de Universidad Virtual de la Universidad de Guadalajara y actualmente está desarrollando una tesis orientada al modelado de evaluación de competencias en entornos e-learning. Es profesora titular de tiempo completo en la Universidad Autónoma de Yucatán desde 2002, donde imparte asignaturas del área de robótica, desarrollo web, teoría computacional y metodología de la investigación. Ha sido responsable y colaboradora de proyectos con financiamiento interno y externo. Colabora con las líneas de investigación aplicación de nuevas tecnologías computacionales y desarrollo de software de aplicación.